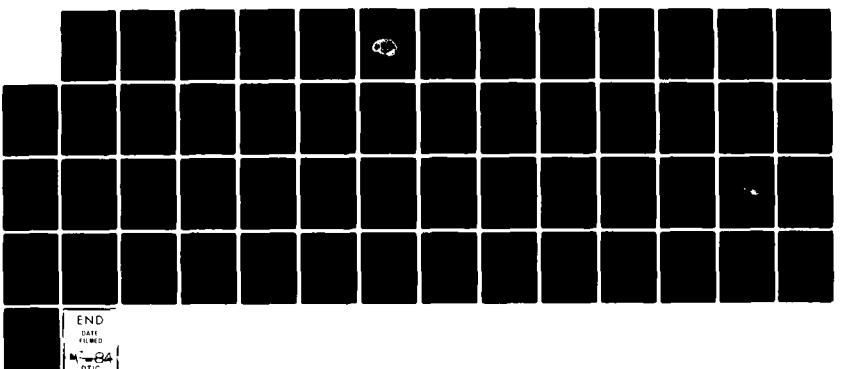
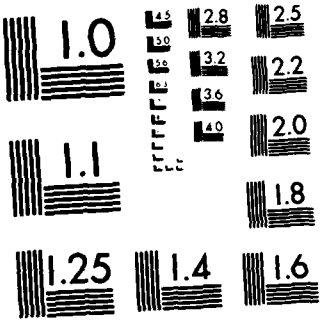
AD-A136 338 NEURON LEARNING IO NETWORK ORGANIZATION(U) BROWN UNIV 1/1
PROVIDENCE RI CENTER FOR NEURAL SCIENCE L N COOPER
20 DEC 83 TR-10 N00014-81-K-0136

UNCLASSIFIED F/G 6/16 NL



END
DATE FILMED
1-84
DTIC



SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

12

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: #10	2. GOVT ACCESSION NO.: AD-A136 338
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Neuron Learning to Network Organization	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Leon N Cooper	8. CONTRACT OR GRANT NUMBER(s): N00014-81-K-0136
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Center for Neural Science, Brown University, Providence, Rhode Island 02912	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 201-484
11. CONTROLLING OFFICE NAME AND ADDRESS: Personnel and Training Research Program, Office of Naval Research, Code 442PT, Arlington, Virginia, 22217	12. REPORT DATE: December 20, 1983 13. NUMBER OF PAGES: 53
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. Publication in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

In Press: Phenomena in Nonlinear Science, North-Holland Publishers.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Neural Networks
Distributed, Associative
Content Addressable Memory
Learning
Visual Cortex
Selectivity
Ocular Dominance

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Progress has been recently made in constructing neural networks that can organize themselves to produce distributed memories. These networks, as well as the proposed procedures by which they modify themselves with experience, are consistent with known neurophysiology as well as with what information may be available at synaptic junctions. The modification assumptions on which these ideas are based have consequences that may be testable in visual cortex. Applied to visual cortex, we assume that between lateral-geniculate and visual cortical cells there exist labile synapses that modify themselves in a fashion consistent with the assump-

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

tions above. Giving the environment an appropriate form, we obtain orientation tuning curves and ocular dominance comparable to what is observed in normally reared adult cats or monkeys. Simulations with binocular input and various types of normal or altered environments show good agreement with the relevant experimental data. Experiments are suggested that could test our theory further.

NEURON LEARNING TO NETWORK ORGANIZATION*

Leon N COOPER

Department of Physics and Center for Neural Science, Brown University, Providence, RI 02912, USA

Introduction

Although we now rank Maxwell among the greatest of 19th century physicists, he wrote that he was adding little to the work Faraday had already done.

> "I have endeavored to make it plain that I am not attempting to establish any physical theory of a science in which I have not made a single experiment worthy of the name, and that the limit of my design is to show how by a strict application of the ideas and methods of Faraday to the motion of an imaginary fluid, everything relating to that motion maybe distinctly represented, and thence to deduce the theory of attractions of electric and magnetic bodies, and of the conduction of electric currents." (*Maxwell, 1856*)

Modesty perhaps, but not entirely unwarranted; for in spite of his enormous talents, the import of his inventions become apparent in the light of later developments with a clarity that, for all of his genius, could have not have been visible to him.

Maxwell's historic achievement was to write down the equations of electricity and magnetism in such a way as to incorporate the experimental discoveries of Coulomb, Ampère and Faraday and to realize that these equations were inconsistent. To make them consistent he was forced to profoundly alter their character, giving rise to a new class of solutions:

*The work on which this article is based was supported in part by the U.S. Office of Naval Research, under contract #N00014-81-K-0136.

propagating waves whose speed (as he calculated using experimental data on electric and magnetic susceptibilities) corresponded very closely to the speed of light.

> "The velocity of transverse undulations in our hypothetical medium, calculated from the electromagnetic experiments of M.M. Kohlrausch and Weber, agrees so exactly with the velocity of light calculated from the optical experiments of M. Fizeau, that we can scarcely avoid the inference that *light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.*" (*Maxwell, 1862.*)

And in a letter to William Thomson (Lord Kelvin):

> "I made out the equations in the country before I had any suspicion of the nearness between the two values of the velocity of propagation of magnetic effects and that of light, so that I think I have reason to believe that the magnetic and luminiferous media are identical." (*Maxwell, 1861.*)

He thus produced a unified field theory of electricity, magnetism and light—the first of its kind. But even this monumental result was just the beginning. For he opened the path to the twentieth century: the Michelson-Morley experiment, relativity, the primacy of field theory and symmetry considerations, Lorentz and, most recently, gauge invariance as general symmetries underlying all physical theories.

This emerges in retrospect. And Maxwell would no doubt be enormously pleased by the great success of the enterprise he began. But he might remind us that his new inventions were preceded by a long exploration of known territory. For most of his working lifetime he applied his physical and mathematical intuition to write down a set of equations that would summarize what was already known. When this could be clearly stated, existing contradictions became apparent-and the new assumptions to remove them relatively quickly made.

Today, I would like to discuss some work that my colleagues and I have been doing recently on the organization of the brain. Although this is somewhat removed from what physicists usually think about, there is a habit of analysis that, I believe, a physicist can profitably bring to complex problems in biology, and perhaps in other areas. Also it is not impossible that a precise understanding of such a complex system could produce surprises—not a new fundamental force or field, but rather a new understanding of the behavior of large interacting systems that could

illuminate the systems of equations that concern us in other domains (as has happened previously in the past generation with the problems of superconductivity, superfluidity and phase transitions).

First, let me attempt a very quick description of the elements of this problem. In Fig. 1, we see a view of the human brain. It is an incredibly complex piece of machinery involving many individual elements—the most relevant of which are known as neurons or nerve cells. It is believed that information processing, memory storage, logical thinking, etc., occurs among the neurons. Neocortex (new cortex)—generally thought to be the thinking part of the brain—is on the surface: this sheet of neurons if spread out is rather large—perhaps several square meters. To fit it into a reasonably sized skull, it had to be folded; typical folds on the surface of the cortex are seen in the Figure. In Fig. 2 a portion of the neural network (in visual cortex) is shown. We see here suggested some of the complexity of the cellular circuitry.

The new part of the brain, cortex, the special biological gift of higher mammals evolved very rapidly, in only a few million years. In contrast, other parts,such as the brainstem that we share with reptiles and that are

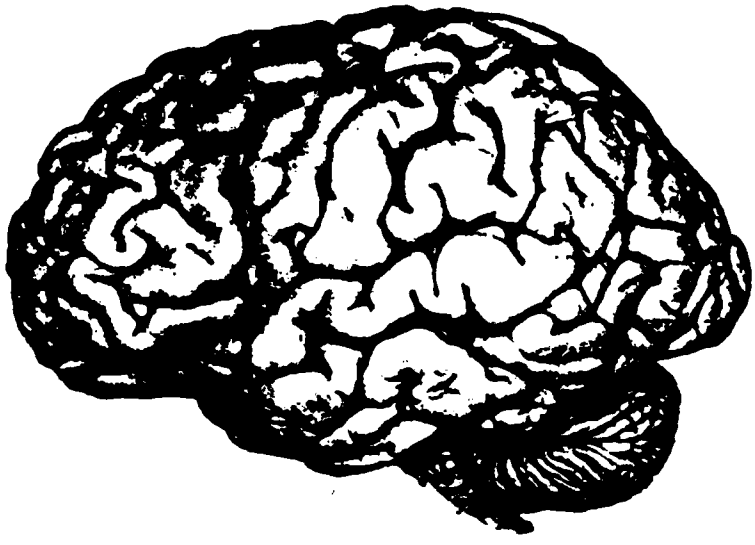
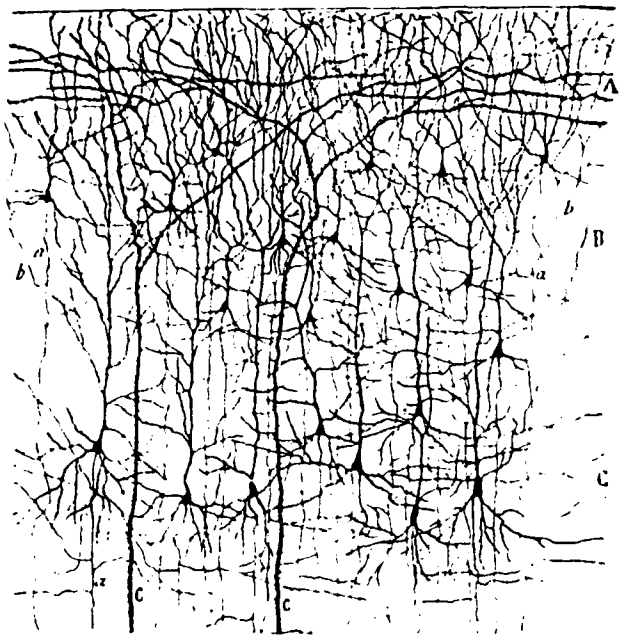


Fig. 1. Side view of the human brain, from DeArmend, Fusco and Dewey, Structure of the Brain.



A, couche plexiforme. — B, couche des petites cellules pyramidales — C, commencement de la couche des cellules pyramidales grandes et moyennes.

Fig. 2. A portion of the neural network in visual cortex, from R. Cajal, Histologie du Systeme Nerveux.

mostly hard-wired and perform a great variety of control functions took hundreds of millions of years to develop. This is a very suggestive fact.

The neurons are a marvellous piece of machinery. Like most cells, they share basic structures to keep themselves alive, but have become extremely specialized. Their primary function is to transmit (and probably also to store) information. The fundamental device utilized by these cells is an excitable membrane. The cell is capable of altering normal ion concentrations in its interior. The proportions of ions such as Na^+, Ca^{++}, K^+ and so on in squid blood are almost those in sea water, which by the way, suggests strongly where the blood comes from (Table 1).

Inside a neuron there is an excess of K^+ and too little Na^+. This is due to a metabolic pump which slowly pumps sodium out and potassium in. (Fig. 3) The pump can be thought of, for practical purposes, as slowly charging a battery. A channel is left open for K^+ ions. Due to the concentration difference, the K^+ ions try to get outside. But since they

Table 1
Concentrations of ions inside and outside freshly isolated axons of squid

Ion	Concentration (mM)		
	Axoplasm	Blood	Seawater
Potassium	400	20	10
Sodium	50	440	460
Chloride	40–150	560	540
Calcium	0.3×10^{-3}*	10	10

*The precise value of ionized intracellular calcium is not known. Data from Hodgkin (1964) and Baker, Hodgkin, and Ridgway (1971), from Kuffler and Nicholls, from Neuron to Brain.

carry positive charge, they build up an electric potential difference across the membrane as they move along the concentration gradient. This potential difference balances the concentration difference when there is a potential difference of about 70 millivolts across the membrane.

Electrical or chemical disturbances on some region of the membrane open the Na^+ channel; there is a rush of ions locally in that region; the negative resting potential, of about −70 millivolts, may jump to a positive potential (about +55 millivolts); this is known as the action potential. Eventually equilibrium is re-established in the original region. But the potential disturbance has spread a bit; this opens the Na^+ channel a bit down the line; then one has another action potential. So the action potential moves down the membrane. Depending on how much depolarization, in a way that I do not have time to go into, one can convert a signal of a given intensity into a given number of action

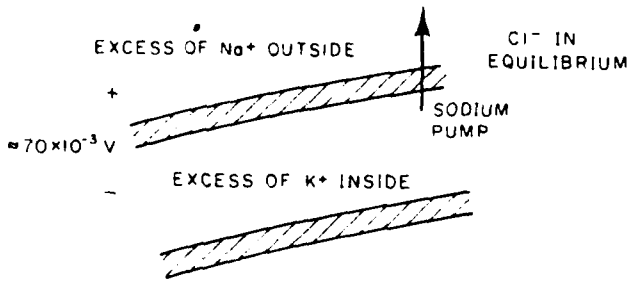


Fig. 3. The neuron membrane.

potentials per second. Thus the information is frequency modulated and can be transmitted with essentially no loss over large distances. It is quite remarkable, considering the materials available, that information can be transmitted with such accuracy.

The transmission of information from one neuron to another is perhaps even more remarkable. An axon of one neuron will terminate, in general, near the dendrite of another in a structure known as a synaptic junction (Fig. 4). Action potentials arriving from the axon initiate release of transmitter substances that diffuse across the synaptic cleft. Upon arrival at the post-synaptic membrane, these transmitters produce changes in membrane conductivity, initiating a flow of ions that alter the dentrite potential. The dendrite potentials propagate to the cell body where they are integrated and determine the firing rate of the post-synaptic cell. Thus the information flow continues.

It is now commonly thought that the synaptic junction may be a means to store information (memory, for example) as well as to transmit it from neuron to neuron. Large networks of neurons connected to other neurons via modifiable synaptic junctions are what we have used to try to construct entities that are capable of holding memory and performing mental acts.

Although the central nervous system contains something of the order of 10–100 billion neurons, it is somewhat depressing to learn that these cells are so specialized that they do not reproduce. Thus when the embryo is given its store of neurons, these are the only ones we will ever have. But cells

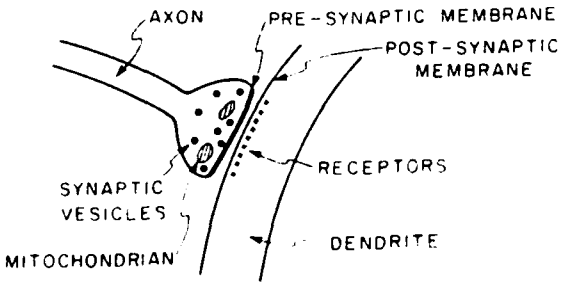


Fig. 4. A synaptic junction

die. We lose an estimated hundred thousand neurons each day so that, at the end of a long lifetime, we have lost a very substantial number of neurons. Any theory that purports to explain the functioning of the central nervous system must explain how it continues to function in spite of the loss of individual neurons.

This is one of the facts that makes the sometime employed analogies between the brain and current computers really very bad. Modern computers perform large numbers of sequential operations very rapidly and very accurately. It is a technological miracle that so many operations can be performed with so few mistakes. The central nervous system works slowly with cycle times that cannot be shorter than a few milliseconds. It is not very accurate: neurons may not fire if they are tired or depressed or just too lazy—in addition, they die.

It may be that current computers can do some of the things that we can do. However, most of the things a computer can do, we do not do very well. We are very poor at arithmetic and sequential logical operations; this is what current computers do very well. We are very good at recognizing things and getting a sense of what is going on. For this computers are very bad. So, if computers can mimic human beings by doing some of things we do, it is very likely that the same thing is being done in a rather different way with rather different hardware.

Although this is a very complex problem, in a certain sense there is much that is known. A set of coupled non-linear differential equations, including time delays, can be written down that in effect summarizes everything that is known about the transmission of electrical signals along excitable membranes and from axon to dendrite in a large coupled and reentrant network of neurons. Such systems can be stationary or can evolve in time by various learning algorithms. Obviously such a set of equations with no simplification is extremely complex and beyond the capacity of current analysis or numerical techniques. The essential point is to make the appropriate approximations and to clearly illuminate the paths connecting assumptions and consequences. Various approaches such as those of Amari (1974), Wilson and Cowan (1973), Edelman (1981), Edelman and Reeke (1982) and Grossberg (1982).

In our work we emphasize the transfer of information between neuron sets to neuron sets; we propose that there is much parallel processing in the central nervous system and that in contrast with machine memory, which is at present local (an event is stored in a specific place) and addressable by locality (requiring some equivalent of indices and files), our memory is distributed and addressable by content or by association.

In addition there need be no clear separation between memory and 'logic', which is a result of association and an out-come of the nature of the memory itself.

Such distributed memories are more like a hologram than a photograph. An individual synaptic junction holds superimposed information concerning many events. In order to obtain a single event, one has to gather information from many junctions. In a system like this, loss of individual neurons decreases the signal to noise ratio but does not lose individual items of information. Therefore, if it is overbuilt in the first place, one can retain a complete memory with an acceptable signal to noise ratio even with loss of neurons.

These ideas as described in more detail in what follows. Although these initial attempts are clearly oversimplifications, our hope is to capture some of the important qualitative features of a very complex phenomenon in a piece of structure that is clear enough so that we can say what follows from what, an explicit enough so that we can make contact with experiment as soon as possible.

1. Theoretical background

1.1 Distributed memory

That most intriguing aspect of memory: its persistence in spite of continual loss of individual neurons over the lifetime of the individual, led us early to the concept of distributed rather than local memory storage. Distributed storage possesses in a very natural way the property of relative invulnerability to the loss of individual storage units. We have been analyzing a class of neural models for the acquisition and storage of distributed momories that display, on a primitive level, features such as recognition, association and generalization, and which suggest some of the mental behavior associated with animal memory and learning (Cooper (1973); Anderson and Cooper (1978)). The mechanisms we employ seem to be plausible biologically and are not inconsistent with known neurophysiology. In addition the networks that results seem to be a reasonable outcome of evolutionary development under the pressure of survival. Some of our ideas are related to or are generalizations of earlier concepts such as perceptrons or similar models (Block, (1962); Block, Knight and Rosenblatt (1962); Minsky and Papert (1969)). In addition holographic or non-local memories have been explored previously (Longuet-Higgins (1968a); Longuet-Higgins (1968b)).

Although the concept of distributed mappings and memory storage is less familiar than those of local storage, distributed mappings and their properties have been discussed (Anderson (1970), (1972); Cooper (1973); Kohonen (1972), (1977)) and probably have already been observed. An example is superior colliculus. In spite of the fact that the retinal afferents that project to the colliculus form a very precise, fine grained map, cells that are just a few millimeters below the very precise cells, respond to stimuli over a wide area of visual space. Thus we have the apparently paradoxical situation—that seems to be true of other parts of the brain as well—that great precision of response is generated by systems composed of cells that progressively show less and less selectivity as the motor output of the system is approached (McIlwain (1976)).

We believe that much of the learning and resulting organization of the central nervous system occurs due to modification of the efficacy or strength of at least some of the synaptic junctions between neurons, thus altering the relation between pre-synaptic and post-synaptic potentials. It is known that small but coherent modifications of large numbers of synaptic junctions can result in distributed memories. Whether and how such synaptic modification occurs, what precise forms it takes, and what the physiological and/or anatomical bases of this modification are, rank among the most interesting questions in this area.

There is direct experimental evidence that at least some modification of synaptic strength occurs in invertebrates (Kandel (1976)) and there are various indications that synaptic modification is a rather general phenomenon (see, for example, (Levy and Steward (1979)). In recent years many conjectures have been made concerning the kind of modification that might occur at synaptic junctions. Kandel and coworkers have shown that modification of the synapse between a sensory and motor neuron of the marine Mollusk Aplysia is the basis of habituation and sensitization. The synaptic modification they have observed can be dependent only on pre-synaptic information. In our work, we have had to assume that synaptic modification is a function of more general variables: local, quasi-local, and global. The presence of quasi-local variables leads to forms of synaptic modification (denoted as Hebbian) that depend on information not immediately available at the synaptic site (e.g., cell firing rates).

These hypotheses have been developed in some detail (Nass and Cooper (1975); Cooper, Liberman and Oja (1979); Bienenstock, Cooper and Munro (1982)) and applied to experimental results that have been obtained in visual cortex by many workers over the last generation as well as to higher level network properties. As will be explained more fully, we have been able to obtain agreement with classical visual cortex experi-

mental results and in addition have suggested new experiments. These theoretical results have been obtained with a minimum of anatomical details and have been primarily concerned with single neurons. It is very likely, however, that interactions between cortical neurons play an important role in cortical function as well, perhaps in selectivity of individual cortical cells (Creutzfeldt et al. (1974); Sillito (1975)). In addition individual synapses are generally either inhibitory or excitatory and not both as some theoretical work has assumed for simplicity. One objective of our current research is to extend our results, taking into account a more realistic anatomy so that a more detailed comparison between theory and experiment can be made.

The theoretical ideas mentioned above have also led to several suggestions for new experiments in visual cortex. In these experiments we attempt to verify predictions concerning the connection between specificity and ocular dominance of cortical cells under various rearing conditions. In addition, we investigate connections between ocular dominance and the variety of visual input allowed to the open eye. We expect the results to give us further information about the detailed mechanism of synaptic modification among cortical cells as well as to enable us to determine various system parameters.

For a distributed memory it is the simultaneous or near simultaneous activities of many different neurons (the result of external or internal stimuli) that is of interest. Thus a large spatially distributed pattern of neuron discharges, each of which might not be very far from spontaneous activity, could contain important, if hard to detect, information. Let us consider the behavior of an idealized neural network (that might be regarded as a model component of a nervous system) to illustrate some of the important features of distributed mappings.

Consider N neurons $1, 2 \ldots N$, each of which has some spontaneous firing rate r_{i0}. (This need not be the same for all of the neurons nor need it be constant in time.) We can then define an N-tuple whose components are the difference between the actual firing rate r_i of the ith neuron and the spontaneous firing rate r_{i0}

$$f_i = r_i - r_{i0}. \tag{1.1}$$

By constructing two such banks of neurons connected to one another (or even by the use of a single bank which feeds signals back to itself), we arrive at a simplified model as illustrated in Fig. 5.

The actual synaptic connections between one neuron and another are

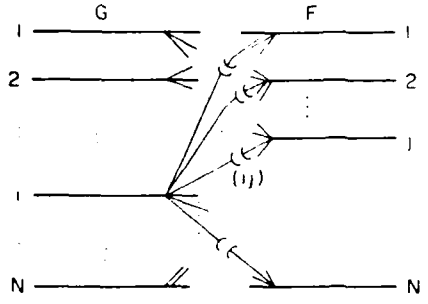


Fig. 5. An ideal distributed mapping. Each of the N input neurons in F is connected to each of the N output neurons in G by a single ideal junction. (Only the connections to i are drawn.)

generally complex and redundant; we have idealized the network by replacing this multiplicity of synapses between axons and dendrites by a single ideal junction which summarizes logically the effect of all of the synaptic contacts between the incoming axon branches from neuron j in the F bank and the dendrites of the outgoing neuron i in the G bank (Fig. 6.). Each of the N incoming neurons, in F, is connected to each of the N outgoing neurons, in G, by a single ideal junction.

Although the firing rate of a neuron depends in a complex and nonlinear fashion on the presynaptic potentials, there is usually a reasonably well defined linear region. Some very interesting network properties are already evident in this linear region. We therefore focus our attention, for the moment, on the region above threshold and below saturation for which the firing rate of neuron i in G, g_i, is mapped from the firing rates of all of the neurons, f_j, in F by:

$$g_i = \sum_{j=1}^{N} A_{ij} f_j . \tag{1.2}$$

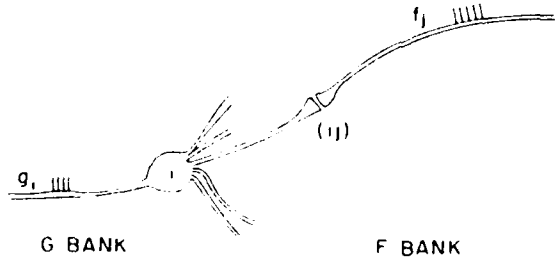


Fig. 6. An ideal synaptic junction

In doing this we are regarding as important average firing rates, and time averages of the instantaneous signals in a neuron (or perhaps a small population of neurons). We are further using the known integrative properties of neurons.

We may then regard $[A_{ij}]$ (the synaptic strengths of the N^2 ideal junctions) as a matrix or a mapping which takes us from a vector in the F space to one in the G space. This maps the neural activities $f = (f_1, f_2 \ldots f_N)$ in the F space into the neural activities $g = (g_1 \ldots g_N)$ in the G space and can be written in the compact form

$$g = Af. \tag{1.3}$$

We propose that it is in modifiable mappings of the type A that human memory is stored. Presently machine memory is local (an event stored in a specific place) and addressable by locality (requiring some equivalent of indices and files). In contrast, human memory is likely to be distributed and addressable by content or by association. In addition for such a memory there need be no clear separation between memory and 'logic'.

It is convenient to write the mapping, A, in the basis of vectors the system has experienced:

$$A = \sum_{\mu\nu} c_{\mu\nu} g^\mu \times f^\nu. \tag{1.4}$$

Here g^μ and f^ν are output and input patterns of neural activity while the $c_{\mu\nu}$ are coefficients reflecting the degree of connection between various inputs and outputs. The symbol, $\times$ represents the 'outer' product between the input and output vectors. Although (1.4) is a well known mathematical form, its meaning as a mapping among neurons deserves some discussion. The ijth element of A gives the strength of the ideal junction between the incoming neuron j in the F bank and the outgoing neuron i in the G bank (Fig. 6.).

Thus, if only f_j is non-zero, g_i, the firing rate of the ith output neuron is

$$g_i = A_{ij} f_j. \tag{1.5}$$

Since

$$A_{ij} = \sum_{\mu\nu} c_{\mu\nu} g_i^\mu f_j^\nu. \tag{1.6}$$

the ijth junction strength is composed of a sum of the entire experience of the system as reflected in firing rates of the neurons connected to this junction. Each experience or association ($\mu\nu$), however, is stored over the entire array of $N \times N$ junctions. This is the essential meaning of a distributed memory: Each event is stored over a large portion of the system, while at any particular local point many events are superimposed.

We show below that the non-local mapping A can serve in a highly precise fashion as a memory that is content addressable and in which 'logic' is a result of association and an outcome of the nature of the memory itself.

1.2. Long and short-term memory

The N^2 junctions, A_{ij}, contain the content of the distributed memory. It could be that a particular junction strength, A_{ij}, is composed of several different components with different lifetimes, e.g.,

$$A = A_{ij}^{(1)} + A_{ij}^{(2)} + \cdots + A_{ij}^{(s)}. \tag{1.7}$$

where the individual $A_{ij}^{(l)}$ might be thought of as corresponding to different physiological or anatomical effects (e.g., changes in numbers of presynaptic vesicles, changes in numbers of postsynaptic receptors, changes in Ca^{++} levels and/or availability, anatomical changes such as might occur in growth or shrinkage of spines). We then have the possibility that the actual memory content (even in the absence of additional learning) will vary with time. For a two-component system we might have

$$A_{ij}^{(t)} = A_{ij}^{(\text{long})}(t) + A_{ij}^{(\text{short})}(t). \tag{1.8}$$

where $A_{ij}^{(t)}$ represents the memory at some time, t, while $A_{ij}^{(\text{long})}$ and $A_{ij}^{(\text{short})}$ have long and short lifetimes. Thus in time $A_{ij}^{(\text{short})}$, will decay, leaving $A_{ij}^{(t)} = A_{ij}^{(\text{long})}$. Whether what is in the short-term memory component is transferred to the long-term component might be determined by some global signal—depending on the interest of the information contained in the short-term component. The existence of such global signals as well as possible anatomical or physiological correlates of short or long-term memory are the subject of some of our current research.

From this point of view the site of long and short-term memory can be essentially identical. At any given time there is a single memory. The distinction between long and short-term memory is contained in the lifetime of the different components of A_{ij}.

1.3. Recognition and recollection

The fundamental problem posed by a distributed memory is the address and accuracy of recall of the stored patterns. Consider first the 'diagonal' portion of A,

$$(A)_{\text{diagonal}} \equiv \mathcal{A} \equiv \sum_{\nu} c_{\nu\nu} g^{\nu} \times f^{\nu} . \tag{1.9}$$

An arbitrary event, e, in the external world mapped by the sensory apparatus into the pattern of neural activity, f, will generate the response in G

$$g = Af .$$

(The pattern, f, might also be the result of some other internal pattern of neural activity.) If we equate recognition with the strength of this response, say the inner product (g, g), then the mapping A will distinguish between those events it contains, the f^{ν}, $\nu = 1, 2, \ldots K$ and other events separated from these.

The work 'separated' in the above context requires definition. Suppose the vectors f^{ν} are thought to be independent of each other, and to satisfy the requirements that, on the average

$$\sum_{i=1}^{N} f_i^{\nu} = 0, \qquad \sum_{i=1}^{N} (f_i^{\nu})^2 = 1 . \tag{1.10}$$

Any two such vectors have components which are random with respect to one another so that a new vector, f, presented in the F bank as above gives a noise like response in the G bank since on the average (f^{ν}, f) is small. The presentation of a vector seen previously, f^{λ}, however, gives the response in the G bank

$$\mathcal{A} f^{\lambda} = c_{\lambda\lambda} g^{\lambda} + \text{noise} . \tag{1.11}$$

It can be shown that if the number of imprinted events, K, is small compared to the dimensionality, N, the signal-to-noise ratios are reasonable.

If we define separated events as those which map into orthogonal vectors, then clearly a recognition matrix composed of K orthogonal vectors $f^1, f^2, \ldots, f^K$:

$$\mathcal{R} = \sum_{\nu=1}^{K} c_{\nu\nu} g^{\nu} \times f^{\nu} \tag{1.12}$$

will distinguish between those vectors contained and all vectors separated from (perpendicular to) these. Further, the response of the system to a vector previously recorded is unique and completely accurate

$$\mathcal{R} f^{\lambda} = c_{\lambda\lambda} g^{\lambda} . \tag{1.13}$$

In this special situation, the distributed memory is as precise as a localized memory.

In addition, this type of memory has the interesting property of recalling an entire associated vector g^{λ} even if only part of f^{λ} is presented. Let

$$f^{\lambda} = f_1^{\lambda} + f_2^{\lambda} . \tag{1.14}$$

If only part of f^{λ}, say f_1^{λ} is presented, we obtain

$$\mathcal{R} f_1^{\lambda} = c_{\lambda\lambda}(f_1^{\lambda}, f_1^{\lambda}) g^{\lambda} + \text{noise} . \tag{1.15}$$

The result is the entire response to the full f^{λ} with a reduced coefficient plus noise.

1.4. Association

If we now take the point of view that presentation of the event e^{ν} which generates the vector f^{ν} is recollected if

$$\mathcal{R} f^{\nu} = c_{\nu\nu} g^{\nu} + \text{noise} . \tag{1.16}$$

Then the off-diagonal terms

$$\mathcal{A} = \sum_{\mu \neq \nu} c_{\mu\nu} g^{\mu} \times f^{\nu} , \tag{1.17}$$

may be interpreted as containing associations between events initially separated from one another.

For such terms the presentation of event e^{ν} will generate not only g^{ν} (which is equivalent to the recollection of e^{ν}) but also, and perhaps more weakly, g^{μ} which should result with the presentation of e^{μ}. Thus, for

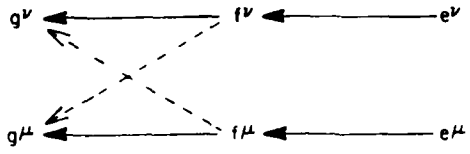


Fig. 7. An ideal association.

example, if g^μ will initiate some response, originally a response to e^μ. The presentation of e^ν when $c_{\mu\nu} \neq 0$ will also initiate this response.

We can thus divide the association matrix A into two parts:

$$A = \sum_{\mu\nu} c_{\mu\nu} g^\mu \times f^\nu = \mathscr{R} + \mathscr{A}, \tag{1.18}$$

where

$$\mathscr{R} = (A)_{\text{diagonal}} \equiv \sum_{\nu} c_{\nu\nu} g^\nu \times f^\nu, \tag{1.19}$$

$$\mathscr{A} = (A)_{\text{off-diagonal}} \equiv \sum_{\mu \neq \nu} c_{\mu\nu} g^\mu \times f^\nu. \tag{1.20}$$

The $c_{\mu\nu}$ are then the direct recollection and association coefficients. Some of the consequences of the properties discussed in the last two sections are the subject of some of our current and continuing research and are further discussed in Subsection 1.5, 1.6, and 1.7.

1.5. Network modification, learning

The properties described above require coherence among many synaptic junctions. We therefore ask: According to what rule and by what means do neurons modify themselves to form a matrix of junctions with the properties of memory? A major effort of our research is to elucidate this question.

Such a modification rule can be cast in the form of stochastic or deterministic differential equations dependent on variables that we classify as local, quasi-local and global.

$$\dot{A}_{ij} = \Phi(g_i, f_j, A_{ij}, t, \ldots). \tag{1.21}$$

Different such rules lead to various types of memories. In the following sections several rules rules for plasticity will be examined. The recollec-

tion-association memory (1.18) described above is obtained from the following simple bilinear modification rule:

$$\delta A_{ij} \quad g_i f_j. \tag{1.22}$$

This δA_{ij} is proportional to the product of the differences between the actual and the spontaneous firing rates in the pre- and post-synaptic neurons i and j. (This is one realization of Hebb's form of synaptic modification (Hebb (1949)).) The addition of such changes to A for all associations $g^\mu \times f^\mu$ results finally in a mapping with the properties discussed in the previous sections.

Synaptic modification dependent on inputs alone, of the type already directly observed in Aplysia (Kandel and Tauc (1965); Castellucci and Kandel (1974)) is sufficient to construct a simple memory—one that distinguishes what has been seen from what has not, but does not easily separate one input from another. To construct a mapping of the form above, however, requires synaptic modification dependent on information that exists at different places on the neuron membrane, what we call two (or higher) point modification.

In order that this take place, information must be communicated from, for example, the axon hillock to the synaptic junction to be modified. This implies the existence of a means of internal communication of information within a neuron—in the above example in a direction opposite to the flow of electrical signals (Cooper (1973)). The junction ij, for example, must have information of the firing rate f_i (which is locally available) as well as the firing rate g_i which is somewhat removed (Fig. 8).

One possibility could be that the integrated electrical signals from the

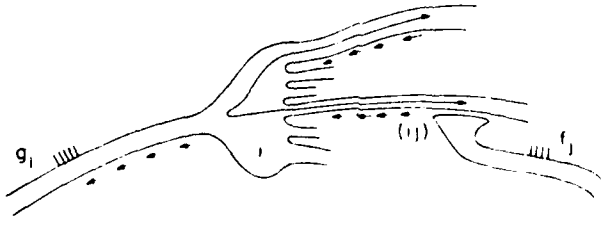


Fig. 8. Two point modification

dendrites produce a chemical or electrical response in the cell body which controls the spiking rate of the axon and at the same time communicates (by backward spiking, for example) to the dendrite ends the information of the integrated slow potential. Another possibility is that dendritic shafts act as somewhat independent units so that the local integrated dendritic potentials interacting with the potentials incoming to the individual spines combine to produce changes in spine shape and resistivity. Such changes might be observable in anatomical studies and are the subject of one of our current research projects.

One might guess that once the physiological mechanism for such communication was available, different types of two (or higher) point modification evolved in various ways. It is tempting to conjecture that a liberating evolutionary step was just the development of this means of internal communication that, coupled with the ability of synapses to modify, created the possibility for a new organization principle.

There is a variety of means by which the coefficient A_{ij} might be modified, given that the necessary information is available at the ijth junction. Among these might be growth of additional or change in electrical properties of dendrite spines, addition of new synaptic junctions, activation of synaptic junctions previously inactive, changes in membrane resistivity and/or changes in the amount of transmitter or receptor in a synapse. Although some structural changes have been observed, there is little evidence yet to choose among the possibilities mentioned above. This is the subject of much current research.

1.6. Passive modification

To make the modification

$$\delta A_{ij} = g^{\alpha}_i \cdot f^{\alpha}_j \,. \tag{1.23}$$

by any of the mechanisms suggested above, the system must have the signal distribution f^{α} in its F bank and g^{α} in its G bank. It is easy to obtain f^{α} since this is mapped from either an external event or is some internal pattern. But to get g^{α} in the G bank is more difficult since this in effect is what the system is trying to learn.

In what we denote as active learning, the system is presented with some f^{α}, searches for a response, and is given some indication of when it is coming closer. When by some procedure or another it finds the 'right' response, say g^{α}, it is 'rewarded' and responds to the reward by printing into A the information:

$$\delta A_{ij} = \eta g_i^{\omega} f_j^{\lambda} . \tag{1.24}$$

(The information is available at the time of the reward since at that time the system is mapping f^{λ}, responding g^{ω}, and thus has just the desired spiking frequencies in the F and G banks of neurons.) Active learning probably describes a type of learning in which a system response to an input is matched against an expected or desired response and judged correct or incorrect.

However, there is a type of learning that does not seem from visible external indications to require this type of a search procedure. It is the type of learning in which, as far as can be seen, an animal is placed in an environment and seems to learn to recognize and to recollect in a far more passive manner.

To arrive at an algorithm which produces what we call passive learning, we utilize a distinction between forming an internal representation of events in the external world as opposed to producing a response to these events that is matched against what is expected or desired in the external world.

The simple but important idea is that *the internal electrical activity that in one mind signals the presence of an external event is not necessarily (or likely to be) the same electrical activity that signals the presence of that same event for another mind.* There is nothing that requires that the same external event be mapped into the same neural patterns by different animals. The event e^{ν} which for one animal is mapped into the signal distributions f^{ν} and g^{ν}, in another animal is mapped into f'^{ν} and g'^{ν}. What is required for eventual agreement between animals in their description of the external world is not that electrical signals mapped be identical but rather that the relation of the signals to each other and to events in the external world be the same (Fig. 9.).

If we now allow the output of a cell to be determined be the input to that cell and the already existing synaptic junction strengths, as well as by possible noise-like fluctuations (making no prior requirement on what the output should be), we arrive at a mathematical formulation of what we call passive modification (Cooper (1973)):

$$\delta A_{ij} = g_i f_j - \sum_{k=1}^{N} A_{ik} f_k f_j . \tag{1.25}$$

It has been shown in the above reference that with a simple form of passive modification a system generates its own response to incoming

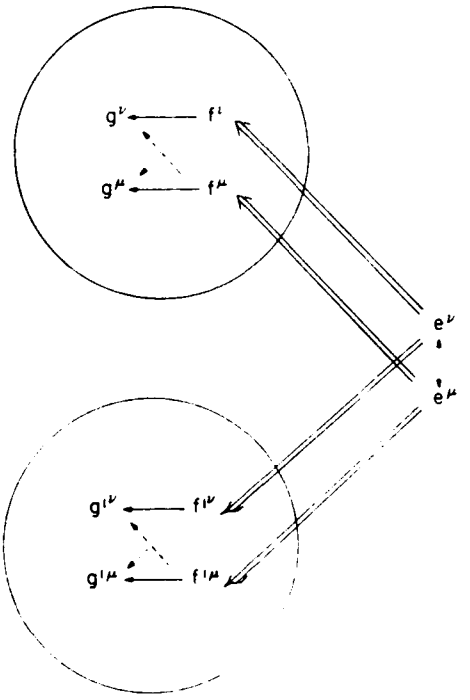


Fig. 9. Representations in two different systems of the same external fabric of events. The two representations are not identical, but they each stand in a one-to-one relation to the external fabric and to each other.

patterns in such a way as to construct distributed mappings that can function as memories capable of recognition and association. To a limited extent these mappings can be regarded as internal representations of what has arrived from the outside world. It has further been shown (Nass and Cooper (1975)) that a form of passive modification can result in the formation of feature detectors or threshold response units which learn to respond to repeated patterns even in the absence of any initial bias. Such units can serve to perform some nonlinear separations.

More detailed discussion of the consequences of these modification procedures and the properties of some of the mappings that result is contained in the references cited above. The application of these ideas to visual cortical cells is discussed in Section 2.

1.7. Feature abstraction

Some networks of neurons must have the ability to extract meaningful information from a broad range of input environments. In the case of sensory input to cortex, for example, the system's range is internally constrained by the response characteristics of the sensory neurons and externally by the nature of the stimulus environment. This stimulus environment depends a great deal upon the nature of the creature's surroundings. Precise statements regarding aspects of environmental structure relevant to mathematical models are given in the next section.

Consider the recognition-association memory (1.18) described above. In actual experience, the events to which the system is exposed are not in general highly separated nor are they independent in a statistical sense. There is no reason, therefore, to expect that all vectors, f^ν, printed into A according to the modification rule (1.25) would be orthogonal or even very far from one another. Rather it seems likely that often large numbers of these vectors would lie close to one another. Under these circumstances, a distributed memory might be 'confused' in the sense that it will respond to new events as if they were old, if the new event is close to an old one. It will 'recognize' and 'associate' events never, in fact, seen or associated before.

The memory will tend to categorize stimuli on the basis of the past history of the system. For example, suppose a number of vectors in the memory are of the form

$$f^\nu = f^0 + n^\nu \tag{1.26}$$

where n^ν varies randomly; f^0 will eventually be recognized more strongly than any particular f^ν actually presented. This, of course, is reminiscent of psychological properties called 'generalization' or 'abstraction'. From such a point of view, generalization grows from the loss of detail of individual instances, a trade-off that seems characteristic of distributed systems.

We have here an explicit realization of feature abstraction. This generalizing quality might be described as the result of a built-in directive for inductive logic. The associative memory by its nature takes the step

$$f^0 + n^1,\ f^0 + n^2, \ldots, f^0 + n^\nu, \ldots \longrightarrow f \tag{1.27}$$

which one perhaps attempts to describe in language as passing from particulars: cat^1, cat^2, cat^3 . . . to general: cat.

How fast this step is taken depends on the parameters of the system. By altering these parameters, it is possible to construct mappings which vary from those which retain all particulars to which they are exposed, to those which lose the particulars and retain only common elements—the central vector of any class.

In addition to 'errors' of recognition, the associative memory also makes 'errors' of association. If, for example, all (or many) of the vectors of the class $\{f^\alpha\}$, defined as a class of vectors not very separated from one another, associate some particular g^β so that the mapping contains terms of the form

$$\sum_{\nu=1}^{K} c_{\beta\nu} g^\beta \times f^\nu, \quad f^\nu \in \{f^\alpha\}, \tag{1.28}$$

with $c_{\beta\nu} \neq 0$ over much of $\nu = 1, 2, \ldots, K$, then the new event e^{K+1} which maps into f^{K+1} also in the class $\{f^\alpha\}$ will not only be recognized, the inner product $(\mathcal{A}f^{K+1}, \mathcal{A}f^{K+1})$ being large, but will also associate g^β, $\mathcal{A}f^{K+1} = cg^\beta + \cdots$ as strongly as any of the vectors $f^1 \ldots f^K$ explicitly contained in (1.28).

If errors of recognition lead to the process described in language as going from particulars to the general, errors of association might be described as going from particulars to a universal: cat^1 meows, cat^2 meows . . . $\rightarrow$ all cats meow.

Whatever efficiacy this inductive process has will depend on the order of the world in which the animal system finds itself. If the world is properly ordered, an animal system that 'jumps to conclusions' in the sense above may be better able to adapt and react to the hazards of its environment and thus survive.

By a sequence of mappings of the form above (or by feeding the output of A back to itself) one obtains a fabric of events and connections that is rich as well as suggestive. One easily sees the possibility of a flow of electrical activity influenced both by internal distributed mappings and the external input. This flow is governed not only by direct association coefficients $c_{\mu\nu}$ (which can be explicitly learned) but also by indirect associations due to the overlapping of the mapped events. One can imagine situations arising in which direct access to an event, or a class of events, has been lost while the existence of this event or class of events in A influences the flow of electrical activity.

One problem in making the identifications suggested above is that such systems tend to form excessively large all-encompassing classes. But

means have been devised to limit the extent of class formation. In fact such mappings can be made to separate classes as well as to unite them (Kohonen (1977); Cooper, Liberman and Oja (1979), (CLO); Bienenstock, Cooper and Munro (1982), (BCM)).

Another problem is a direct consequence of the assumption of the linearity of the system. Any state is generally a superposition of various vectors. Thus one has to find a means by which events—or the entities into which they are mapped—are distinguished from one another.

There are various possibilities; neurons are so non-linear that it is not at all difficult to imagine non-linear or threshold devices that would separate one vector from another. Such separation processes compliment generalization processes in that they bring out the differences in an input environment while generalizing cells tune to the component most common to the constituent stimuli. But the occurrence of a vector in a distributed memory in a set of signals over a large number of neurons each of which is far from threshold. A basic problem, therefore, is how to associate the threshold of a single cell or a group of cells with such a distributed signal. One way this might come about has been shown by Nass and Cooper (1975). Another possibility is the stochastic process recently discussed by Hopfield (1982).

In addition to the appearance of 'pontifical' cells or groups of cells, there will be a certain separation of mapped signals due to actual localization of the areas in which these signals occur. For example, optical and auditory signals are subjected to much processing before they actually meet in cortex. It is possible to imagine that identification of optical or auditory signals (as optical or auditory) occurs first from where they appear and their immediate cluster associations. Connections between an optical and an auditory event might occur as suggested in Fig. 10. Although the systems described above are relatively primitive, they

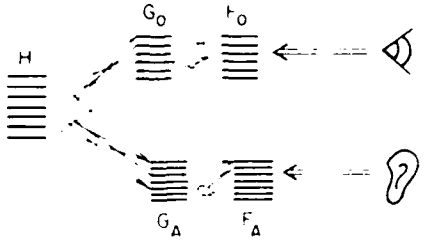


Fig. 10. A model optical auditory system

suggest various psychological properties and are used in our research to construct models of some aspects of behavior and language learning.

2. Application to visual cortex: Comparison of theory with experiment

2.1. Summary of related visual cortex experimental data

The discussion above leads to a central issue: what is the principle of local organization that, acting in a large network, can produce the observed complex behavior of higher mental processes. There is no need to assume that such a mechanism—believed to involve synaptic modification—operates in exactly the same manner in all portions of the nervous system or in all animals. However, one would hope that certain fundamental similarities exist so that a detailed analysis of the properties of this mechanism in one preparation would lead to some conclusions that are generally applicable. We are interested in visual cortex because the vast amount of experimental work done in this area of the brain—particularly area 17 of cat and monkey—strongly indicate that one is observing a process of synaptic modification dependent of the information locally and globally available to the cortical cells.

Experimental work of the last generation, beginning with the pathbreaking work of Hubel and Wiesel (1959, 1962), has shown that there exists cells in visual cortex (areas 17, 18, and 19) of the adult cat that respond in a precise and highly tuned fashion to external patterns, in particular bars or edges of given orientation and moving in a given direction. Much further work (Blakemore and Cooper (1970); Blakemore and Mitchell (1973); Hirsch and Spinelli (1971); Pettigrew and Freeman (1973)) has been taken to indicate that the number and response characteristics of such cortical cells can be modified. It has been observed in particular (Imbert and Buisseret (1975); Blakemore and Van Sluyters (1975); Buisseret and Imbert (1976); and Fregnac and Imbert (1977, 1978)), that the relative number of cortical cells that are highly specific in their response to visual patterns varies in a very striking way with the visual experience of the animal during the critical period.

Most kittens first open their eyes at the end of the first week after birth. It is not easy to assess whether or not orientation selective cells exist at that time in striate cortex: few cells are visually responsive, and the response's main characteristics are generally 'sluggishness' and fatigability. However, it is quite generally agreed that as soon as cortical cells are reliably visually stimulated (e.g., at 2 weeks), some are orientation selec-

tive, whatever the previous visual experience of the animal (cf. Hubel and Wiesel (1963); Blakemore and Van Sluyters (1975); Buisseret and Imbert (1976); Fregnac and Imbert (1978)).

Orientation selectivity develops and extends to all visual cells in area 17 if the animal is reared, and behaves freely, in a normal visual environment (NR): complete "specification" and normal binocularity (about 80% of responsive cells) are reached at about 6 weeks of age (Fregnac and Imbert (1978)). However, if the animal is reared in total darkness from birth to the age of 6 weeks (DR), none or few orientation selective cells are then recorded (from 0 to 15%, depending on the authors and the classification criteria); however, the distribution of ocular dominance seems unaffected (Blakemore and Mitchell (1973); Imbert and Buisseret (1975); Blakemore and Van Sluyters (1975); Buisseret and Imbert (1976); Leventhal and Hirsch (1980); Fregnac and Imbert (1978)). In animals whose eyelids have been sutured at birth, and which are thus binocularly deprived of pattern vision (BD), a somewhat higher proportion (from 12 to 50%) of the visually excitable cells are still orientation selective at 6 weeks (and even beyond 24 months of age) and the proportion of binocular cells is less than normal (Wiesel and Hubel (1965); Blakemore and Van Sluyters (1975); Kratz and Spear (1976); Leventhal and Hirsch (1977); Watkins, et al., (1978)).

Imbert and Buisseret have classified cortical cells that respond to visual stimuli into three groups—aspecific, immature, and specific. They, Fregnac and Imbert have measured the relative proportions of these groups depending on the visual experience of the animal. The distribution of the different cell types in three age groups is shown in Fig. 11.

Examination of these results, which were obtained from the study of 1050 cells, shows that cells having some of the highly specific response properties of adult visual cortical neurons, especially concerning orientation selectivity are present in the earliest stages of post-natal development independent of visual experience (Fregnac and Imbert (1977, 1978)). However, visual experience between 17 and 70 days is critical in determining the evolution of these cells. Animals reared normally showed a marked increase in the number of specific cells as compared with aspecific. (The period between 17 and 28 days is usually sufficient to reach the normal adult level of specificity.) The reverse is true for animals reared in the dark. A statistical analysis of this evolution, performed by Fregnac (1978) shows clearly the striking dependence of the ratio of sharply tuned to broadly tuned cells depending on the experience of the animal.

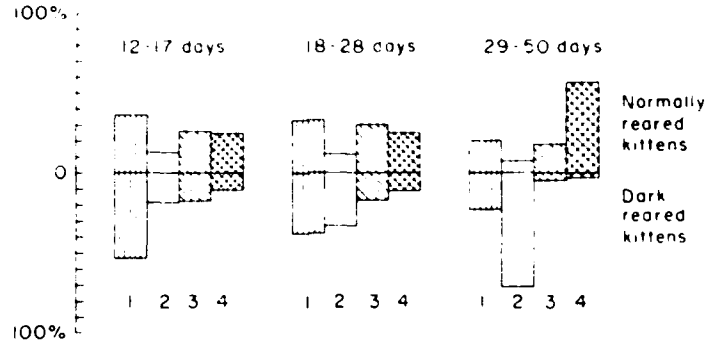


Fig. 11. Distribution of the different types of cells in three age groups in normally reared kittens (upper part) and in dark-reared kittens (lower part). The ordinate is normalized so that the heights are the percentages of cells in the various function groups. Type 1, nonactivable (vert. stripes); 2, nonspecific (open); 3, immature (diagonal stripes); 4, specific (cross-hatched). (From Fregnac and Imbert (1977, 1978).)

In addition, as has been shown by Imbert and Buisseret (1975), Buisseret and Imbert (1976) and Buisseret et al. (1978) as little as six hours of normal visual experience at about 42 days of age can alter in a striking fashion the ratio of specific or immature to aspecific cells (Fig. 12.). That such a short visual experience can change the tuning ratios so markedly is clear evidence of the great plasticity of these cortical cells at the height of the critical period.

Of all visual deprivation paradigms, putting one eye in a competitive advantage over the other has probably the most striking consequences. If monocular lid-suture (MD) is performed during a 'critical' period (ranging from about 3 weeks to about 12 weeks), there is a rapid loss of binocularity to the profit of the open eye (Wiesel and Hubel (1963, 1965)). At this stage, opening the closed eye and closing the experienced one may result in a complete reversal of ocular dominance (Blakemore and Van Sluyters (1974)). A disruption of binocularity that does not favor one of the eyes may be obtained, for example, by provoking an artificial strabismus (Hubel and Wiesel (1965)) or by an alternating monocular occlusion, which gives both eyes an equal amount of visual stimulation (Blakemore (1976)). In what follows, we call this uncorrelated rearing (UR).

These results seem to us to provide direct evidence for the modifiability of the response of single cells in the cortex of a higher mammal according to its visual experience. Depending on whether or not patterned visual

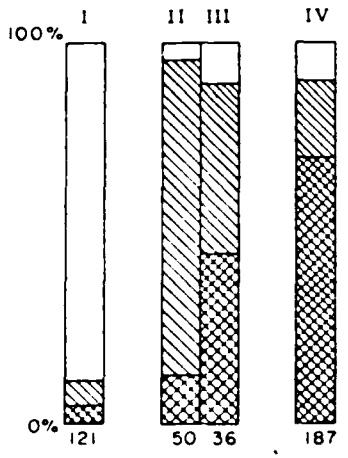


Fig. 12. Distribution in percentage of the three types of visual cortical units (area 17) recorded after 6 hours of visual exposure for 6-week-old dark-reared kittens. Columns: I, dark-reared kittens; IV, normally reared kittens. During 8 hours of exposure, conditions were: in II and III, freely moving; in III, 12 hours in the dark followed the 6 hours of exposure. Numbers of visual cells recorded are given under each column. Specific cells (cross-hatched) are activated by oriented stimuli within a sharp angle (<60°). Immature cells (diagonal stripes) are activated by oriented stimuli within a larger angle (<150°). Nonspecific cells (open) are activated by nonoriented stimuli moving in any direction. A statistical analysis reveals no significant difference in the percentage of immature and specific units between columns III and IV. Therefore it may be that for a 6-week-old dark-reared kitten, a 6-hour exposure to visual input followed by 12 hours in the dark is sufficient to produce a distribution of cortical cells similar to that of normally reared animals. (From Buisseret et al. (1978).)

information is part of the animal's experience, the specificity of the response of cortical neurons varies widely. Specificity increases with normal patterned experience. Deprived of normal patterned information (dark-reared or lid-sutured at birth, for example) specificity decreases. Further, even a short exposure to patterned information after six weeks of dark-rearing can reverse the loss of specificity and produce an almost normal distribution of cells.

We do not claim and it is not necessary that all neurons in visual cortex be so modifiable. Nor is it necessary that modifiable neurons are especially important in producing the architecture of visual cortex. It is our hope that the general form of modifiability we require to construct distributed mappings manifests itself for at least some cells of visual cortex that are accessible to experiment. We thus make the conservative assumption that biological mechanisms, once established, will manifest themselves in more or less similar forms in different regions. If this is the

case, modifiable individual neurons in visual cortex can provide evidence for such modification more generally.

2.2. Modification of cortical synapses: global and local variables

To apply the general theoretical ideas of the previous section to visual cortex, we introduce the following notation. Consider a cortical cell as shown in Fig. 13:

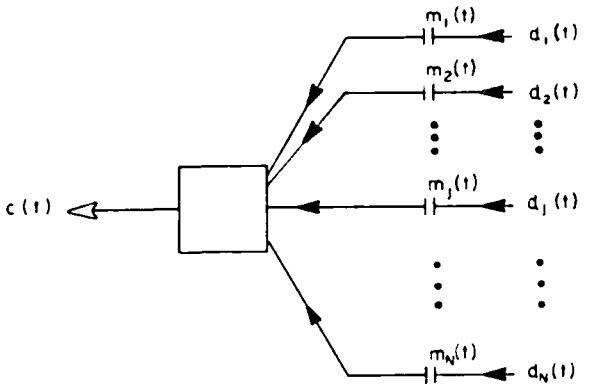


Fig. 13. A model neuron which processes the input $d(t)$ according to the synaptic weights $m(t)$ to yield the response $c(t)$.

Replacing equations (1.1) and (1.2) we write

$$c(t) = \sum_j m_j(t) d_j(t) , \tag{2.1}$$

where $c(t)$ is the output at time t, $m_j(t)$ is the efficacy of the jth synapse at time t, $d_j(t)$ is the jth component of the input at time t (the firing frequency of the jth presynaptic neuron) and Σ_j denotes summation over j, i.e., over all presynaptic neurons. We can then write:

$$\begin{aligned} m(t) &= (m_1(t), m_2(t), \ldots, m_N(t)) , \\ d(t) &= (d_1(t), d_2(t), \ldots, d_N(t)) , \\ c(t) &= m(t) \cdot d(t) . \end{aligned} \tag{2.2}$$

$m(t)$ and $d(t)$ are real-valued vectors, of the same dimension, N, i.e., the

number of ideal synapses onto the neuron, and $c(t)$ is the inner product (or 'dot product') of $m(t)$ and $d(t)$. The vector of synaptic efficacies at time t, $m(t)$, is called the *state* of the neuron at time t. (Note that $c(t)$ as well as all components of $d(t)$ represent firing frequencies that are measured from the level of average spontaneous activity; thus they might take negative as well as positive values; $m_j(t)$ is dimensionless.)

We can now formulate the question: What is the local principle of organization, by asking what is the change in time of $m_j(t)$ (the jth synapse onto the cortical cell, receiving inputs $d_j(t)$) and on what variables does this depend.

The various factors that influence synaptic modification may be divided broadly into two classes—those dependent on global and those dependent on local information. Presumably, global information in the form of chemical or electrical signalling influences most (or all) modifiable junctions of a given type in a given area in the same way. Evidence for the existence of global factors that affect development may, for instance, be found in Kasamatsu and Pettigrew (1976, 1979), Singer (1979, 1980) and Buisseret et al. (1978), Baer and Daniels (1983) and Bear et al. (1983). On the other hand, local information available at each modifiable synapse can influence each junction in a different manner.

An early proposal as to how local information could affect synaptic modification was made by Hebb (1949). His, now classical, principle was suggested as a possible neurophysiological basis for operant conditioning: "when an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased." Thus the increase of the synaptic strength connecting A to B is dependent upon the correlated firing of A and B. Such a correlation principle has inspired the work of many theoreticians on various topics related to learning, associative memory, pattern recognition, organization of neural mappings (retinotopic projections) and development of selectivity of cortical neurons.

It is fairly clear that in order to actually use Hebb's principle one must state conditions for synaptic decrease as specific as those for synaptic increase: if synapses are allowed only to increase, all synapses will eventually saturate; no information will be stored and no selectivity will develop (see for example Scinowski, (1977a,b)). What is required is thus a complementary statement to Hebb's principle giving conditions for synaptic decrease. Such a statement is given in what follows.

For a general form of synaptic modification, we write:

$$\dot{m}_i = F(d_i \ldots m_i; d_k \ldots \bar{c} \ldots c; X \ldots Y \ldots Z), \tag{2.3}$$

where the first set of variables $d_i \ldots m_i$ are what we call local, the second set of variables $d_k \ldots \bar{c}$ what we call quasi-local, while the third set what we call global. Local variables such as $d_i \ldots m_i$ are those directly at the synaptic site. Thus any information would be directly available. Quasi-local variables are those such as $d_k \ldots c, \bar{c}$. These are physically connected to the synaptic site by the cell itself. However, in order that the information they contain be available, some means of internal cellular communication must be assumed. Note that we include among these such variables as $\bar{c}$ (the averaged activity of the cell over time). Global variables are called $X \ldots Y \ldots Z$.

In work done in the past few years we have explored a form of synaptic modification that can be written as follows. Referring to the jth synaptic junction:

$$\dot{m}_i = \phi(c, \bar{c})d_i - \varepsilon m_i. \tag{2.4}$$

Note that as in passive modification, the output of a cell is determined by the input and the already existing synaptic strengths as well as by noise-like fluctuations. The precise form of ϕ is not critical as long as it has certain general characteristics. Cooper, Liberman, and Oja (1979), (CLO) showed that if the function ϕ goes through zero then the sharpness of the tuning curve is altered by the visual experience of the animal in agreement with what is observed. This modification might be called 'Hebbian' when the output is above the modification threshold, θ_M, and 'anti-Hebbian' when the output is below this threshold. The function, ϕ, is also assumed to have a dependence on global variables, not explicitly written. CLO thus assumed that the modifiability of a synaptic junction is dependent on events that occur at different parts of the same cell and on the rate at which the cell responds. They proved several theorems which show that with this form of passive modification there is an increase in the specificity of the response of a cortical cell to visual input (sharpening of its tuning curve) when that cell is exposed to stimuli that are the result of normal patterned visual experience and a loss of specificity when that cell is exposed to noise-like input, such as might be expected when an animal is dark-reared or raised with eyelids sutured. Specificity can be regained, however, with a return of input due to patterned vision.

In addition to this basic behavior, simulations and mathematical results on the asymptotic states of the neural network show some more subtle

phenomena that depend upon values of system parameters. Of note are the rate of decay (forgetting per unit time), the strength of selective modification of synaptic junctions, the interaction of modifiable with non-modifiable synapses, and the different statistical properties of noise factors.

The reason for the increase of selectivity is the crossover of the ϕ function from the negative to the positive region at the modification threshold θ_M. This was recognized by CLO to be associated with some property of the cell, possibly the average firing rate. This idea was enlarged and extended by Bienenstock, Cooper and Munro (1982) (BCM) and applied to a great variety of situations in visual cortex. The essential idea of BCM was to allow ϑ_M to vary non-linearly with the average activity of the cell, $\bar{c}$. Doing this they achieved a variety of desirable properties as well as a theoretical structure in excellent agreement with available experimental data. The crucial point in the choice of the function $\phi\ (c, \bar{c})$ is the determination of the threshold $\theta_M(t)$, i.e., the value of c at which $\phi(c, \bar{c})$ changes sign. A candidate for $\theta_M(t)$ is the average value of the postsynaptic firing rate, $\bar{c}(t)$. The time average is meant to be taken over a period T preceding t much longer than the membrane time-constant τ so that $\bar{c}(t)$ evolves on a much slower timescale than $c(t)$. This can usually be approximated by averaging over the distribution of inputs for a given state $m(t)$

$$\bar{c}(t) = m(t) \cdot \bar{d}\,. \tag{2.5}$$

This results in an essential feature, the *instability of low selectivity points*. (This can be most easily seen at zero selectivity equilibrium points, where, with any perturbation, the state is driven away from this equilibrium, whatever the input.)

Therefore, if stable equilibrium points exist in the state space, they are of high selectivity. However, do such points exist at all? The answer is generally yes provided that the state is *bounded from the origin and from infinity*. These conditions, instability of low-selectivity equilibria as well as boundedness, are fulfilled by a single function $\phi(c, \bar{c})$ if we define $\theta_M(t)$ to behave as a *nonlinear function* of $\bar{c}(t)$, for example, a power. The exponent should then be larger than 1. The final requirement on $\phi\ (c, \bar{c})$ thus reads:

$$\text{sign}\ \phi(c, \bar{c}) = \text{sign}\left(c - \left(\frac{\bar{c}}{c_0}\right)^p \bar{c}\right) \quad \text{for } c > 0\,, \tag{2.6}$$

$$\phi(0, \bar{c}) = 0 \quad \text{for all } \bar{c}\,,$$

where c_0 and p are two fixed positive constants. The threshold $\theta_M(\bar{c}) = (\bar{c}/c_0)^p\bar{c}$ thus serves two purposes: allowing threshold modifications when $\bar{c} = c_0$ as well as driving the state from regions such that $\bar{c} \ll c_0$ or $\bar{c} \gg c_0$. The process of synaptic growth, starting near 0 to eventually end in a stable selective state, may be described as follows. Initially, $\bar{c} \ll c_0$ hence $\phi(c, \bar{c}) > 0$ for all inputs in the environment: the responses to all inputs grow. With this growth $\bar{c}$ increases, thus increasing θ_M. Now some inputs result in postsynaptic responses that exceed θ_M, while others—those whose direction is far away (close to orthogonal) from the favored inputs—give a response less than θ_M. The response to the former continues to grow while the response to the latter decays. This results in a form of *competition between incoming patterns* rather than competition between synapses. The response to unfavored patterns decays until it reaches 0, where it stabilizes, for $\phi(0, \bar{c}) = 0$ for any $\bar{c}$. The response to favored patterns grows until the mean response $\bar{c}$ is high enough, and the state stabilizes. This occurs in spite of the fact that many complicated geometrical relationships may exist between different patterns, i.e., that they are not orthogonal since different patterns may and certainly do share common synapses.

Any function, ϕ, that satisfies (2.6) will give these qualitative results. The precise form of this function (e.g., the numerical values of p and c_0) will affect the detailed behavior of the system such as rate of convergence, height of the maximum for a selective cell as well as a variety of other more subtle effects. We are presently investigating the consequences of various detailed assumptions concerning the form of $\phi(c, \bar{c})$ and comparing these with existing and proposed experiments. In doing this we hope to arrive at a detailed understanding of the form of the function that controls synaptic modification.

We note also that with this form of modification, the control of θ_M by a global signal (in addition to $\bar{c}$) could produce the following results: If θ_M is set to be very large the cell's response would diminish. This will result in a behavior that is like that described by Eric Kandel in Aplysia habituation experiments. If θ_M is set very low the cell will rapidly increase its response to a stimulus. This could be related to a type of sensitization in which the sensitizing signal has the effect of resetting θ_M to a very low level. For a variable θ_M as will be shown below applied to visual cortex, one gets increasing and decreasing of selectivity such as those seen in experimental results over the last generation. We thus have the possibility that a single mechanism of modification, functioning in slightly different ways can account for a variety of experimental data in both invertebrates and vertebrates.

values in the space of inputs to the neuron $\mathcal{N}$. The variable d represents a random input to the neuron, and is characterized by its probability distribution that may be discrete or continuous. (During normal development, the input to the neuron [or neural network] is presumably distributed uniformly over all orientations. In abnormal rearing conditions [e.g., dark reared] the input during development could be different from the input for measuring selectivity. How this should be translated in the formal space R^N will be discussed later.) This distribution defines an environment, mathematically a random variable d. Selectivity is estimated (before, or *after* development) with respect to this same environment. Obviously, $\mathrm{Sel}_d(\mathcal{N})$ always falls between 0 and 1, and the higher selectivity of $\mathcal{N}$ in d, the closer $\mathrm{Sel}_d(\mathcal{N})$ is to 1.

We analyze the behavior of (2.4) for $\varepsilon = 0$. The behavior depends critically on the environment, that is, on the distribution of the stationary stochastic process, d. Two classes of distributions may be considered:

(a) *Discrete distributions* (K possible inputs $d^1, \ldots, d^K$): These are generally assumed to occur with the same probability $1/K$. The process d is then a jump process which randomly assumes new values at each time increment. The vector m is (roughly) a Markov process.

(b) *Continuous distributions*: in work of BCM, the only continuous distribution that is considered is a uniform distribution d over a closed 1-parameter curve in the input space R^N. Although the principles underlying the convergence to selective states are intuitively fairly simple, mathematical analysis of the system is not entirely straightforward, even for the simplest d. Mathematical results, obtained only for certain discrete distributions, are of two types: (1) equilibrium points are locally stable if and only if they are of highest available selectivity with respect to the given distribution of d, (2) given any initial value of m in the state space, the probability that $m(t)$ converges to one of the maximum selectivity fixed points as t goes to infinity is 1. Results of the second type are much stronger, and require a tedious geometrical analysis. Results are stated here in a somewhat simplified form. For exact statements and proofs, the reader is referred to Bienenstock (1980) or to BCM (1982). To illustrate, we study the simple case where d takes on values on only two possible input vectors d^1 and d^2, that occur with the same probability and let $\varepsilon = 0$ in (2.4):

$$P[d = d^1] = P[d = d^2] = 1/2 .$$

Whatever the real dimension N of the system it reduces to two dimen-

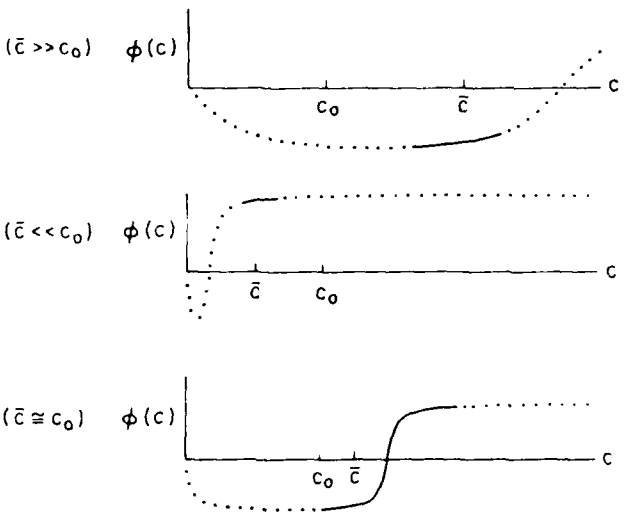


Fig. 14. A function satisfying condition (2.6). The three diagrams show the behavior of $\phi(c, \bar{c})$ as a function of c for three different constant values of $\bar{c}$. In each diagram, the *solid part* of the *curve* represents $\phi(c, \bar{c})$ in the vicinity of $\bar{c}$, which is the relevant part of this function.

2.3. *Some mathematical results*

Selectivity

It is common usage to estimate the orientation selectivity of a single visual cortical neuron by measuring the half-width and half-height—or an equivalent quantity—of its orientation tuning curve. The selectivity is then measured with respect to a parameter of the stimulation, namely the orientation, which takes on values over an interval of 180°. In our work, various kinds of inputs are considered, e.g., formal inputs with a parameter taking values on a finite set of points, rather than a continuous interval. It will then be useful to have a convenient general index of selectivity, defined in all cases. We propose the following:

$$\mathrm{Sel}_d(\mathcal{N}) = 1 - \frac{\text{mean response of } \mathcal{N} \text{ with respect to } d}{\text{maximum response of } \mathcal{N} \text{ with respect to } d}. \qquad (2.7)$$

With this definition, selectivity is estimated *with respect to*, or *in* an *environment for the neuron*, that is, a random variable d that takes on

sions. (Any component of m outside the linear subspace spanned by d^1 and d^2 will eventually decay to 0 due to the uniform decay term.)

Analytic results in two dimensions

It follows immediately from the definition that the maximum value of $\mathrm{Sel}_d(m)$ in the state space is 1/2. It is reached for states m which give null response when d^1 comes in (i.e., are orthogonal to d^1) but positive response for d^2—or vice versa. Minimum selectivity, namely 0, is obtained for states m such that $m \cdot d^1 = m \cdot d^2$. Equilibrium states of both kinds indeed exist.

Lemma 1. *Let d^1 and d^2 be linearly independent and d satisfy $P[d = d^1] = P[d = d^2] = 1/2$. Then for any ϕ satisfying* (2.6) *the system* (2.4) *admits exactly 4 fixed points, m^0, m^1, m^2, and $m^{1,2}$ with: $\mathrm{Sel}_d(m^0) = \mathrm{Sel}_d(m^{1,2}) = 0$, and $\mathrm{Sel}_d(m^1) = \mathrm{Sel}_d(m^2) = 1/2$. (Here the superscripts indicate which of the d are not orthogonal to m. [m^0 is the origin.] Thus for instance $m^1 \cdot d^1 > 0$, $m^1 \cdot d^2 = 0$.)*

The behavior of the system depends on the geometry of the inputs, in the present case on $\cos(d^1, d^2)$. The crucial assumption that is needed here is that $\cos(d^1, d^2) > 0$. This is a reasonable assumption which is obviously satisfied if all components of the inputs are positive, as is assumed in some models (Von der Malsburg (1973); Perez et al. (1975)). We may then state the following:

Theorem 1. *Assume that in addition to the conditions of Lemma 1, $\cos(d^1, d^2) > 0$. Then m^0 and $m^{1,2}$ are unstable, m^1 and m^2 are stable, and whatever its initial value, the state of the system converges almost surely (i.e., with probability 1) either to m^1 or to m^2.*

Theorem 1 is the basic result in the 2-dimensional setting: it characterizes evolution schemes based on *competition between patterns*, saying that the state eventually reaches maximal selectivity even when the two input vectors are very close to one another. Obviously this requires that some of the synaptic strengths be negative since the neuron has linear integrative power. Inhibitory connections are thus necessary to obtain selectivity. Some selectivity is also realizable with no inhibitory connections—not even 'intracortical' ones— if the integrative power is appropriately nonlinear. However, whatever the nonlinearity of the in-

tegrative power, Theorem 1 could not hold for evolution equations based on *competition between converging afferents.*

In Theorem 1, we have a discrete sensory environment which consists of exactly two different stimuli—a situation, although simple mathematically, not often encountered in nature. It may, however, very well correspond to a visual environment restricted to only horizontally and vertically oriented contours, present with equal probability. Theorem 1 then predicts that cortical cells will develop a selective response to one of the two orientations, with no preference for either (other than what may result from initial connectivity). Thus, on a large sample of cortical cells, one should expect as many cells tuned to the horizontal orientation as to the vertical one. So far, no assumption is made on intracortical circuitry. We discuss this later.

The proof of Theorem 1 is based on the existence of *trap regions* around each of the selective fixed points:

Theorem 2. *Under the same conditions as in Theorem* 1, *there exists around* $m^1(m^2)$ *a region* $F^1(F^2)$, *such that once the state enters* $F^1(F^2)$, *it converges almost surely to* $m^1(m^2)$.

The meaning of Theorem 2 is the following: once $m(t)$ has reached a certain selectivity, it cannot 'switch' to another selective region. Applied to cortical cells in a patterned visual environment, this means that once they become sufficiently committed to certain orientations, they will remain committed to those orientations (provided that the visual environment does not change), becoming more selective as they stabilize to some maximal selectivity. Theorems 1 and 2 are illustrated in Fig. 15.

It is worth mentioning that when $\cos(d^1, d^2) < 0$, the situation is much more complicated: trap regions don't necessarily exist and periodic asymptotic behavior, i.e., limit cycles, may occur, bifurcating from the stable fixed points when $\cos(d^1, d^2)$ becomes too negative (see Bienenstock (1980)).

Higher dimensions

We now turn to the case where d takes on K values. The following is easily obtained:

Lemma 2. *Let* $d^1, d^2, \ldots, d^K$ *be linearly independent and* d *satisfy* $P[d = d^1] = \cdots = P[d = d^K] = 1/K$. *Then, for any* ϕ *satisfying* (2.6), (2.4) *admits exactly* 2^K *fixed points with selectivities* $0, 1/K, 2/K, \ldots, (K-1)/K$. *There are* K *fixed points* $m^1, \ldots, m^K$ *of selectivity* $(K-1)/K$.

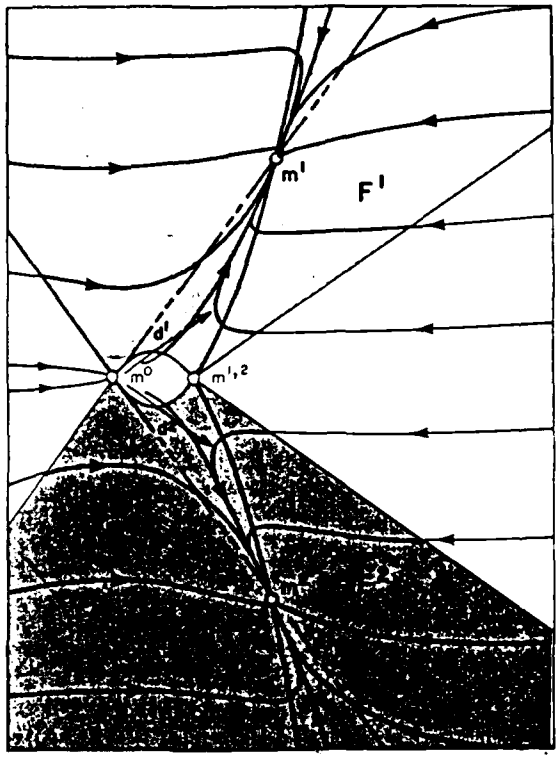


Fig. 15. The phase portrait of equation (2.4) subject to condition (2.6). The diagram shows the trajectories of the state of the neuron starting from different initial points. The final state of the system (m^1 or m^2) is determined when the trajectory enters the corresponding 'trap' (shaded) region (F^1 or F^2).

Obviously, $(K-1)/K$ is also the maximum possible selectivity with respect to d. It means a positive response for one and only one of the inputs. The situation is now much more complicated than what it was with only 2 inputs: it is not obvious whether in all cases assuming that all the cosines between inputs are positive is sufficient to yield stability of the maximum selectivity fixed points. However, we may state the following:

Theorem 3. *Assume, in addition to the conditions of Lemma 2, that* $d^1, \ldots, d^K$ *are all mutually orthogonal or close to orthogonal. Then the K fixed points of maximum selectivity are stable, and, whatever its initial value, the state of the system converges almost surely to one of them.*

The proof of Theorem 3 also involves trap regions around the K maximally selective fixed points, and the analog of Theorem 2 is true here.

Although the general case has not yet been solved analytically, as will be seen later, computer simulations suggest that for a fairly broad range of environments if $d^i \cdot d^j > 0$, even if $d^1, \ldots, d^K$ are far from being mutually orthogonal, the K fixed points of maximum selectivity are stable.

Simulations suggest further that even if the $d^1, \ldots, d^K$ are *not* linearly independent and are very far from being mutually orthogonal, the asymptotic selectivity is close to its maximum value with respect to d.

Analytic results in two dimensions and computer simulations in higher dimensions indicate that the form of synaptic modification described here leads, in general, to the evolution of maximum selectivity with respect to the environment. We are trying to extend the linear analysis of stability performed in two dimensions to higher dimensions. Stability analysis has been attempted on systems of K dimensions for general linearly independent environment (Cooper et al. (1982). The same arguments that lead to statements of stability in two dimensions apply in this general case. However, the technical difficulty increases. The problem may be stated in terms of a Kth order eigenvalue equation. Local stability for an $s = 1$ fixed point will be assured if the eigenvalues of the matrix of coefficients of the K differential equations are negative. Similarly, the instability of points for which $s > 1$ would be characterized by the presence of positive eigenvalues. Since this matrix of coefficients exhibits some symmetry, there is hope that the problem could be solved analytically (for reasonable size K, the system of equations could be solved numerically for special cases). This kind of analytic statement would confirm that the states of high selectivity observed in computer simulations are indeed stable asymptotic states.

The monocular problem: A simple circular environment

We now apply this theory to the problem of orientation selectivity and binocular interaction in primary visual cortex. The ordinary development of these properties in mammals depends to a large extent on normal functioning of the visual system (i.e., normal visual experience) during the first few weeks or months of postnatal life. This has been demonstrated many times by various experiments, based mainly on the paradigm of

rearing the animal in a restricted sensory environment. We show that the theory described above can account for both normal development and development in restricted visual environments.

Consider first a classical test environment used to construct the tuning curve of cortical neurons. This environment consists of an elongated light bar successively presented or moved in all orientations—in a random sequence—in the neuron's receptive field. Thus all the parameters of the stimulus are constant except one, the orientation, which is uniformly distributed on a circularly symmetric closed path. We assume that the retino-cortical pathways map this family of stimuli to the cortical neuron's space of inputs in such a way as to preserve the circular symmetry (as defined below). Thus, the typical theoretical environment that will be used for constructing the neuron's tuning curve is a random variable d uniformly distributed on a circularly symmetric closed one-parameter family of points in the space R^N. The parameter coding orientation in the receptive field is, in principle, continuous. However, for the purpose of numerical simulations, the distribution is made discrete. Thus, d takes on values on the points $d_1, \ldots, d^K$.

The requirement of circular symmetry is expressed mathematically as follows: the matrix of inner products of the vectors $d^1, \ldots, d^K$ is circular (*i.e., each row is obtained from its nearest upper neighbor* by shifting it one column to the right) and the rows of the matrix are unimodal. A random variable, d, uniformly distributed on such a set of points will be, hereafter, called a *circular environment*. Such a d may be roughly characterized by 3 parameters: N, K and a measure of the mutual geometrical closeness of the d's, for instance the minimum value of $\cos(d^1, d^i)$ over the environment.

We are now faced with the difficult problem of specifying the stationary stochastic process that represents the time-sequence of inputs to the neuron during development. To begin, we simplify the problem by giving the stochastic process exactly the same distribution as the circular d defined above. In doing so, we assume that development of orientation selectivity is to a large extent independent of other parameters of the stimulus, e.g., contrast, shape, position in the receptive field, retinal disparity for binocular neurons, etc. The elementary stimulus for a cortical neuron is a rectilinear contrast edge or bar. Any additional pattern present at the same time in the receptive field is regarded as random noise. (A discussion of this point is given in Cooper et al. (1979)).

Simulations show the following behavior:

(1) The state converges rapdily to a fixed point, or *attractor*.

(2) Various such attractors exist. For a given d and ϕ they all have the same selectivity, which is close to its maximum value in d.

(3) The asymptotic tuning curve is always unimodal. One may thus talk of the preferred orientation of an attractor.

(4) There exists an attractor in each possible orientation.

(5) If there is no initial preference, all orientations have equal probability of attracting the state. (Which one will become favored depends on the exact sequence of inputs). This does not hold for environments which are not perfectly circular, at least for a single neuron system as the one studied here.

The system thus behaves exactly as expected from the results of the preceding section.

The binocular problem: a more complex environment

We now consider a binocularly driven cell. The firing rate of the neuron at time t is now given by

$$c(t) = m_l(t) \cdot d_l(t) + m_r(t) \cdot d_r(t) . \tag{2.8}$$

with evolution schemes for 'left' and 'right' states m_l and m_r straightforward generalizations of (2.4). We have partitioned the input vector space into a left space and a right space; hence m goes to (m_l, m_r) and d becomes (d_l, d_r). Since d_l and d_r can be independent, the topology of the environment is potentially more complex.

Various possibilities exist for the input (d_l, d_r): one may wish to consider normal rearing (both d_l and d_r circular and presumably highly correlated), monocular deprivation, binocular deprivation, and so on. The vector (d_l, d_r) is a stationary stochastic process, whose distribution is one of the following, depending on the experimental situation one wishes to reproduce:

Normal Rearing (NR):

$d_l(t) = d_r(t)$ for all t, and d_l is circular. (Noise terms that may be added to the inputs may or may not be stochastically independent.)

Uncorrelated Rearing (UR):

d_l and d_r are i.i.d. (independent identically distributed): they have the same circular distribution, but no statistical relationship exists between them.

Binocular Deprivation (BD):

The $2N$ components of (d_l, d_r) are i.i.d.: d_l and d_r are uncorrelated noise terms.

Monocular Deprivation (MD):

d_l is circular, d_r is a noise term: $d_r = n$.

In the NR case, the inputs from the two eyes to a binocular cell are probably well correlated. We therefore assume that they are equal, which is mathematically equivalent. The BD distribution represents dark discharge.

Uncorrelated or strabismic rearing (UR) involves presenting fully two independent circular environments (a 'toroidal' environment). The final state can be either monocular and specific or binocular and specific with no correlation between the patterns perferred by the two eyes.

The results of binocular deprivation or (correlated) normal rearing are just those of the monocular case. We assume that binocular stimuli presented during NR are exactly correlated so that each pattern incident to the left-eye synapses is consistently accompanied by a corresponding pattern to the right-eye synapses. Since the left and right components of each pair are identical, the cell tunes to the same pattern in each eye. Binocularly deprived input environments consisted of stimulus components uniformly distributed over some range with zero mean. In this case (BD), the average response of the cell is null and so ϕ is always non-negative, resulting in random fluctuations of the synaptic state.

The development of a neuron receiving patterned input from only one eye (and uniform noise from the other) is somewhat surprising. The response curve goes to maximum selectivity with respect to the open eye, but, consistent with observation, the response to the closed eye does *not* fluctuate randomly. Rather the neuron becomes nonresponsive to inputs to the deprived eye. Asymptotic convergence to this state is assured *regardless of the initial state*. The theoretical implications for the reverse suture (RS) paradigm are straightforward: A monocularly deprived neuron, having reached a monocular selective state is driven to another monocular selective state preferring the *newly opened eye* upon reversal of suture.

This behavior relies upon some activity, albeit purely random, to be present in the afferents from the closed eye. Such noise may be due to diffuse light through the eyelid or spontaneous firing of LGN and/or retinal neurons. As a neuron becomes selective with respect to the open eye, patterns which are preferred give a response near threshold whereas the other patterns give a much lower response. In either case ϕ is near zero. Noise accompanying a preferred pattern drives the neuron from the modification threshold, so the deprived synapses grow stronger. However, the opposite effect weakens the synapses when non-preferred patterns are presented. A mathematical demonstration of this argument, given in Appendix C of Bienenstock et al. (1982), is presented in 2.4.

2.4. Comparison of theory with classical experimental results

The simulated behavior of neurons in visual cortex with binocular connectivity is illustrated in Fig. 16. The seemingly inconsistent experimental results (MD vs. BD) are faithfully reproduced by computer simulation. Each of these paradigms was tested in both deterministic and stochastic simulation algorithms over several pattern sets. The model withstood considerable noisy input; indeed successful simulation of some paradigms (RS in particular) *required* that a noiselike component accompany the 'pure' inputs.

Simulations of the behavior of the system in these different environments give the following:

NR: all asymptotic states are selective, binocular and have matching preferred orientations for stimulation through each eye.

BD: the motion of the state (m_l, m_r) resembles a random walk. (The small exponential decay term is necessary here in order to prevent large fluctuations.) The two tuning curves therefore undergo random fluctuations that are essentially determined by the second-order statistics of the input d. As can be seen from the figure, these fluctuations may sometimes result in a weak orientation preference or unbalanced ocular dominance. However, the system never stays in such states very long; its average state on the long run is perfectly binocular and nonoriented. Moreover, whatever the second-order statistics of d and the circular environment in which tuning curves are assessed, a regular unimodal orientation tuning curve is rarely observed, and selectivity never exceeds 0.6. We may thus conclude that orientation selectivity as observed in the NR case (both experimental and theoretical) cannot be obtained from purely random synaptic weights. It is worth mentioning here that prolonged dark rearing has been reported to increase response variability (Leventhal and Hirsch (1980)); a similar observation was made by Fregnac and Bienenstock (1981).

MD and RS: The only stable equilibrium points are monocular and selective. The system converges to such states whatever the initial conditions. In particular, this accounts for reverse suture experiments (Blakemore and Van Sluyters (1974); Movshon (1976)).

UR: In contrast to NR, monocular as well as binocular equilibria exist. The asymptotic state generally observed with $m_l(0) = m_r(0) = 0$ is monocular. (This should be attributed to the mismatched inputs from the two eyes, as is done by most authors.) Asymptotic states are selective, and when they are binocular, preferred orientations through each eye do not necessarily coincide. It should be mentioned here that Blakemore and

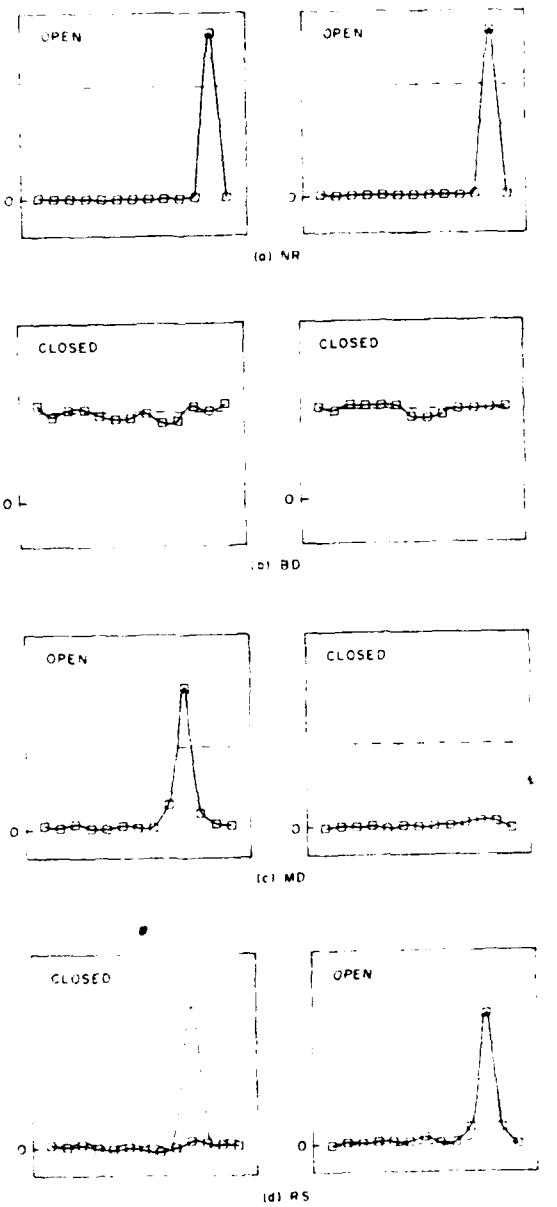


Fig. 16. Computer simulations of various rearing conditions. Initial (dashed) and final (solid) responses to the two eyes are shown separately (left/right).

Van Sluyters (1974) report that after a period of alternating monocular occlusion, the remaining binocular cells may differ in their preferred orientations for stimulation through each eye.

Selectivity and ocular dominance

As an example of the kind of new and subtle effects that are contained in this theory, we consider in detail the sequence in which ocular dominance and selectivity develop in the monocularly deprived environment.

According to (2.8) the firing rate of a binocularly driven neuron at time t is given by

$$c(t) = m_l(t) \cdot d_l(t) \cdot d_r(t).$$

In a situation corresponding to monocular deprivation—patterned information to one eye (right), noise to the other (left)—we can write for the environment

$$d = (d_r, n),$$

and for the set of synaptic weights

$$m = (m_r, m_l).$$

Where m_r and m_l are the synaptic weights from the right and left eyes respectively.

In this situation m_r goes to one of its selective fixed points as in the monocular case. The only fixed point for m_l in the noise-like environment is zero; but this is unstable in the monocular case. It is instructive to follow the behavior of m_l in this binocular case.

Let (x_r, x_l) be a small perturbation from equilibrium. The motion at point $(m_r^* + x_r, x_l)$ is given by:

$$\dot{x}_r = \phi(m_r^* \cdot d_r + x_r \cdot d_r + x_l \cdot n, m_r^* \cdot \bar{d}_r + x_r \cdot \bar{d}_r) d_r, \tag{2.9r}$$

$$\dot{x}_l = \phi(m_r^* \cdot d_r + x_r \cdot d_r + x_l \cdot n, m_r^* \cdot \bar{d}_r + x_r \cdot \bar{d}_r) n, \tag{2.9l}$$

where we assume that the noise has zero mean.

We analyze separately, somewhat informally, the behavior of the two equations. The stability of (2.9r) is immediate from the stability of the selective state m_r^* in the circular environment d_r. To analyze (2.9l) we divide the range of the right eye input d_r into three classes:

(i) d_r is such that $m_r^* \cdot d_r$ is either far above threshold, θ_M, and therefore $\phi(m_r^* \cdot d_r, m_r^* \cdot \bar{d}_r) > 0$, or far below threshold, θ_M, (but still positive) and therefore $\phi(m_r^* \cdot d_r, m_r^* \cdot \bar{d}_r) < 0$;

(ii) d_r is such that $m_r \cdot d_r$ is near threshold, θ_M, and therefore $\phi(m_r^* \cdot d_r, m_r^* \cdot \bar{d}_r) \simeq 0$;

(iii) d_r is such that $m_r^* \cdot d_r \simeq 0$ and again $\phi(m_r^* \cdot d_r, m_r^* \cdot \bar{d}_r) \simeq 0$.

For the first class of inputs, the sign of ϕ is determined by d_r alone, hence 2.9*l* is the equation of a random walk. To investigate the behavior of 2.9*l* in the two other cases, we neglect the term x_r and linearize ϕ around the relevant one of its two zeros. It is easy to see that case (ii) yields

$$\dot{x}_l \simeq \varepsilon_1 (x_l \cdot n) n \,, \tag{2.10}$$

whereas in case (iii) one obtains

$$\dot{x}_l \simeq -\varepsilon_2 (x_l \cdot n) n \,, \tag{2.11}$$

where ε_1 and ε_2 are *positive* constants, measuring respectively the absolute value of the slope of ϕ at the modification threshold and at zero.

Since n is a noise-like term, its distribution is presumably symmetric with respect to x_l so that averaging (2.10) and (2.11) yields respectively

$$\dot{x}_l \simeq \varepsilon_1 \bar{n}_0^2 x_l \,, \tag{2.12}$$

$$\dot{x}_l \simeq -\varepsilon_2 \bar{n}_0^2 x_l \,, \tag{2.13}$$

where $\bar{n}_0^2$ is the average squared magnitude of the noise input to a single synaptic junction from the closed eye.

We thus see that input vectors from the first class move x_l randomly, inputs from the second class drive it away from 0, whereas inputs from the third drive it toward 0. In the case where the range of d_r is a set of K linearly independent vectors and m_r^* is of maximum selectivity, $(K-1)/K$, case (i) does not occur at all. (The random contribution occurs only before the synaptic strengths from the open eye have settled to one of their fixed points.) Case (ii) occurs only for one input, say d_r^1, with $m_r^* \cdot d_r^1$ exactly equal to threshold, θ_M, and (iii) occurs for the other $K-1$ vectors which are all orthogonal to m_r^*. In the general case (d_r any circular environment), the more selective m_r^* with respect to d_r, the higher the proportion of inputs belonging to class (iii), the class that yields (2.13) i.e., that brings x_l back to 0.

The stability of the global system still depends on the ratio of the quantities ε_1 and ε_2 as well as on the statistics of the noise term n (e.g. its mean square norm). We may however formulate two general conclusions. First, under reasonable assumptions (ε_1 of the order of ε_2 and the mean square norm of n of the same order as that of d_i) $x_i = 0$ is stable on the average for a selective m_i^*. Second, the residual fluctuation of x_i around 0, essentially due to inputs d_i in classes (i) and (ii), is smaller for highly selective m_i^*'s than it is for mildly selective ones.

Thus, one should expect that in a monocularly deprived environment nonselective neurons tend to remain binocularly driven. In addition since it is the non-preferred inputs from the open eye accompanied by noise from the closed eye (case three) that drive the response to the closed eye to zero, if inputs to the open eye were restricted to preferred inputs (case two) even a selective cell would remain less monocular.

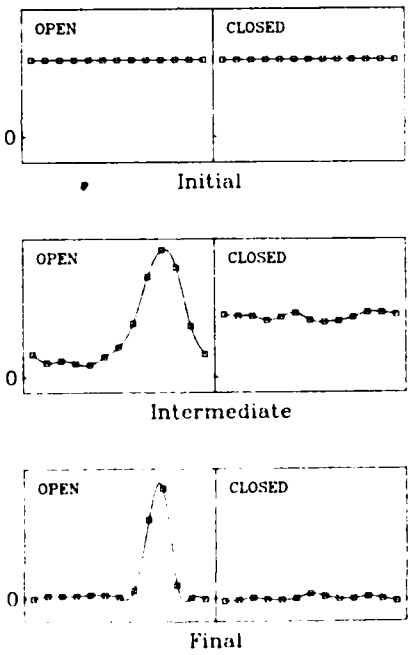


Fig. 17. Progression of development of selectivity and ocular dominance. Note that selectivity develops for the open eye *before* the response to the closed eye is driven to zero.

To bètter confront these ideas with experiment, the single (BCM) neuron must be placed in a network with the anatomical features of visual cortex, a network in which inhibitory and excitatory cells receive input from LGN and from each other. This has been done (Scofield and Cooper to be published). Their conclusions are similar to those above with explicit further statements concerning the independent effects of excitatory and inhibitory neurons on selectivity and ocular dominance. For example, shutting off inhibitory cells lessens selectivity and alters ocular dominance giving 'masked synapse' effects.

Quantitative tests of progressions such as those shown in Figure 17 are in progress in our laboratory. We hope that such experiments can provide detailed comparisons with theory and provide us with a sensitive tool for determining synaptic modification among various classes of neurons--a possible entry to the process by which the nervous system organizes itself.

REFERENCES

Amari, S.I. (1974), "Among Theoretical Theory of Nervenets", Adv. in Biophysics 6, 75-120.

Anderson, J. A. (1970), "Two models for memory organization using interacting traces", Math. Biosciences 8, 137-160.

Anderson, J. A. (1972), "Simple neural network generating and interactive memory", Math. Biosciences 14, 197-220.

Anderson, J. A. and L. N Cooper (1978), "Biological organization of memory", Encyclopaedia Universalis France S. A., (Pluriscience), pp. 168-175.

Bear, M. F., M. A. Paradiso, M. Schwartz, S. B. Nelson, K. M. Carnes, and J. Daniels, "Two Methods of Catecholamine Depletion in Kitten Visual Cortex Yield Different Affects on Plasticity," Nature 302:245247 (1983).

Bear, M. F. and J. D. Daniels, "The Plastic Response to Monocular Deprivation Persists on Kitten Visual Cortex after Chronic Depletion of Norepinephrine," J. Neurosci. 3:407-416 (1983).

Bienenstock, E. "A Theory of Development of Neuronal Selectivity, Ph.D. Thesis, Division of Applied Mathematics and Center for Neural Science. L. N Cooper Thesis Supervisor. (1980).

Bienenstock, E. L., L. N Cooper, and P. W. Munro, "Theory for te development of neuron selectivity: Orientation selectivity and binocular interaction in visual cortex", Jour. of Neurosci. 2, 32-48 (1982).

Blakemore, C. and G. F. Cooper, "Development of the brain depends on the visual environment", Nature 228, 477-478 (1970).

Blakemore, C. and D. F. Mitchell, "Environmental modification of the visual cortex and the neural basis of learning and memory", Nature 241, 467 (1973).

Blakemore, C. and R. C. Van Sluyters, "Reversal of the physiological effects of monocular deprivation in kittens. Further evidence for a sensitive period", J. Physiol. London 237, 195-216 (1974).

Blakemore, C. and R. C. Van Sluyters, "Innate and environmental factors in the development of the kitten's visual cortex", J. Physiol. London 248, 663-716 (1975).

Blakemore, C. Van Sluyters, R. C. and Moyshon, J. A., "Synaptic competition n the kitten's visual cortex", Cold Spring Harbor Symp. Quant. Biol. 40, The Synapse, 601-609 (1975).

Blakemore, C. "The conditions required for the maintenance of binocularity in the kitten's visual cortex", J. Physiol. 261, 423-444 (1976).

Block, H. D. , "The perception: A model for brain functioning I.", Rev. Mod. Phys. 34, 123-135 (1962).

Block, H. D., Bw. Knight, Jr., and F. Rosenblatt, "Analysis of a four-layer series coupled perception, II.", Rev. Mod. Phys. 34, 135-142 (1962).

Buisseret, P. and M. Imbert, "Visual cortical cells. Their developmental properties in normal and dark reared kittens", J. Physiol. London 255, 511-525 (1976).

Buisseret, P. E. Gary-Bobo, and M. Imbert, "Ocular motility and recovery of orientational properties of visual cortex neurons in dark reared kittens", Nature 272, 816-817 (1978).

Castellucci, V. F. and E. R. Kandel, "A quantal analysis of the synaptic depression underlying habituation of the gill-withdrawal reflex in Aplysia", Proc. Nat. Acad. Sci. 71, 5004 (1974).

Cooper, L. N, "A possible organization of animal memory and learning", in Proc. Nobel Symposium on Collective Properties of Physical Systtems, ed. by B. Kindquist and S. Lindquist, (Academic Press, New York) pp. 252-264 (1974).

Cooper, L. N, F. Liberman and E. Oja, "A theory for the acquisition and loss of neuron specificity in visual cortex", Biol. Cybernetics 33, 9-28 (1979).

Creutzfeldt, O.D., U. Kuhnt, and L. A. Benevento, "An intracellular analysis of visual cortical neurons to moving stimuli: Responses in a cooperative neuronal network", Exp. Brain Res. 21, 251 (1974).

Edelman, G. M., "Group Selection as the Basis for Higher Brain Functions" in The Organization of the Cerebral Cortex, ed. by F. O. Schmitt, F. G. Worden, G. Adelman, S. G. Dennis, pp. 535-563, MIT Press, Cambridge, Mass. (1981).

Edelman, G. M. and G. N. Reeke, Jr., "Selective Networks Capabe of Representative Transformations, Limited Generalizations and Associative Memory," PNAS Biol. 79, pp. 2091-2095 (1982).

Fregnac, Y. and M. Imbert, "Cinetique de developpment du cortex visuel", J. Physiol. Paris 6, Vol. 73 (1977).

Fregnac, Y. "Cinetique de developpment du cortex visual primarie chez le chat. Effets de la privation visuelle binoculaire et modele de maturation de la selective a l'orientation, Doctoral thesis, University Rene Descartes (1978).

Fregnac, Y. and M. Imbert, "Early development of visual cells in normal and dark-reared kittens: Relationship between orientation selectivity and ocular dominance", J. Physiol. London 278, 27-44 (1978).

Fregnac, Y. and E. Bienenstock, "Specific functional modifications of individual cortical neurons, triggered by vision and passive eye movement in immobilizedk kittens", in Pathophysiology of the Visual System: Documentta Ophthalmologica, ed. by L. Maffei, Vol. 30, pp. 100-108, Dr. W. Junk, The Hague (1981).

Grossberg, S., "Studies of Mind and Brain", D. Reidel Dordrecht, Holland; Boston, USA; London, England (1982).

Hebb, D. O., "The Organization of Behavior, (Wiley, New York) p. 62 (1949).

Hirsch, H.V.B. and D. N. Spinelli, "Modification of the distribution of receptive field orientation in cats by selective visual exposure during development", Exp. Brain Res. 12, 509-527 (1971).

Hopfield, J. J., "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", PNAS, Vol. 79, pp. 2554-2558 (1982).

Hubel, D. H. and T. N. Wiesel, "Receptive fields of single neurons in the cat striate cortex", J. Physiol. London 148, 574-591 (1959).

Hubel, D. H. and T. N. Wiesel, "Receptive fields, binocular interactions and functional architecture in the cat's visual cortex", J. Physiol. London 160, 106-154 (1962).

Hubel, D. H. and T. N. Wiesel, "Receptive fields of cells in striate cortex of very young, visually inexperienced kittens", J. Neurophysiol. 26, 994-1002 (1963).

Hubel, D. H. and T. N. Wiesel, "Binocular interaction in striate cortex of kittens reared with artificial squint", J. Neurophysiol. 28: 1041-1059 (1965).

Imbert, M. and Y. Buisseret, "Receptive field characteristics and plastic properties of visual cortical cells in kittens reared with or without visual experience", Exp. Brain Res. 22, 2-36 (1975).

Kandel, E. R. and L. Tauc, "Mechanism of heterosynaptic facilitation in the giant cell of the abdominal ganglion of Aplysia depilans", J. Physiol. 181, 28 (1965).

Kandel, E. R., "Cellular Basis of Behavior: An Introduction to Behavioral Neurobiology, (W. H. Freeman, San Francisco) (1976).

Kasamatsu, T. and J. D. Pettigrew, "Depletion of brain catecholamines: failure of ocular dominance shift after monocular occlusion in kittens", Science 194, 206-209 (1976).

Kasamatsu, T. and J. D. Pettigrew. "Preservation of binocularity after monocular deprivation in the striate cortex of kittens treated with 6-hydroxydopamine", J. Comp. Neurol. 185, 139-162 (1979).

Kohonen, T., "Correlation matrix memories", IEEE Trans. on Computers, C. 21, 353-359 (1972).

Kohonen, T., "Associative Memory: A System Theoretic Approach (Springer, Berlin) (1977).

Kratz, K. E. and P. D. Spear, "Effects of visual deprivation and alterations in binocular competition on responses of striate cortex neurons in the cat", J. Comp. Neurol. 170, 141 (1976).

Leventhal, A. G. and H. V. B. Hirsch, "Effects of Early Experience Upon Orientation Selectivity and Binocularity of Neurons in Visual Cortex of Cats", Proc. Natl. Acad. Sci., U.S.A. 74:1272-1276 (1977).

Leventhal, A.G. and H. V. B. Hirsch, "Receptive field properties of different classes of neurons in visual cortex of normal and dark-reared cats", J. Neurophysiol. 43, 1111 (1980).

Levy, W. B. and O. Steward, "Synapses as associative memory elements in the hippocampal formation", Brain Res. 175, 233-245 (1979).

Longuet-Higgins, H.C., "Holographic model of temporal recall", Nature 217, 104 (1968).

Longuet-Higgins, H. C., "The non-local storage of temporal information", Proc. R. Soc. London (Biol.) 171, 327-334 (1968).

Maxwell, J. C., "On Faraday's Lines of Force", Part 1 Transactions of the Cambridge Philosophical Society 10, 27-83 (1856).

Maxwell, J. C., "Letter to William Thomson (Lord Kelvin), (Dec. 10, 1861) (1861).

Maxwell, J. C., "On Physical Lines of Force", Part 3 Philosophical Magazine, (Jan. Feb. 1862), Proposition 16 (1862).

McIlwain, J. T., "Large receptive fields and spatial transformations in the visual system", in Internatonal Review of Physiology: Neurophysiology II., Vol. 10, ed. by R. Porter (University Park Press, Baltimore) (1976).

Minsky, M. and S. Papert, "An Introduction to Computational Geometry", (MIT Press, Cambridge) (1969).

Nass, M. M. and L. N Cooper, "A theory of the development of feature detecting cells in visual cortex", Biol. Cybernetics 19, 1-18 (1975).

Perez, R. L. Glass and R. J. Shaler, "Development of specificity in the cat visual cortes", J. Math. Biol. 1.275 (1975).

Pettigrew, J. D. and R. D. Ferman, "Visual experience without lines: effects on developming cortical neurons", Science 182, 599-601 (1973).

Scofield, C. L. and Cooper, L. N, "Selectivity and Ocular Dominance in Visual Cortex: A Network Theory" (to be published).

Sejnowski, T. J., "Storing covariance will nonlinearly interacting neurons", J. Math. Biol. 4, 303 (1977a).

Sejnowski, T. J., "Statistical constrainst on synaptic plasticity", J. Theor. Biol. 69, 385 (1977b).

Sillito, A. M., "The contribution of inhibitory mechanisms to the receptive field properties of neurons in the cat's striate cortex", J. Physiol. 250, 304-330 (1975).

Singer, W., "Central-core control of visual-cortex functions", in The Neurosciences: Fourth Study Program, ed. by F. Schmit and F. Worden, (MIT Press, Cambridge, Massachusetts), pp. 1093-1109 (1979).

Von der Malsburg, C., "Self-organization of orientation sensitive cells in the striate cortex", Kybernetik 14, 85 (1973).

Watkins, D. W., J. R. Wilson, and S. M. Sherman, "Receptive field properties of neurons in binocular and monocular segments of striate cortex in cats raised with binocular lid suture", J. Neurophysiol. 41, 322 (1978).

Wiesel, T. N. and D. H. Hubel, "Single-cell responses in striate cortex of kittens deprived of vision in one eye", J. Neurophysiol. 26, 1003-1017 (1963).

Wiesel, T. N. and D. H. Hubel, "Comparisons of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens", J. Neurophysiol. 28, 1029-1040 (1965).

Wilson, H. R. and J. D. Cowan, "A Mathematical Theory for the Functional Dynamics of Cortical and Thalamic Nervous Tissues", Kybernetik 13:55-80 (1973).

Wood, C., "Variations on a theme by Lashley: Lesion experiments on the neural models of Anderson, Silverstein, Ritz, and Jones", Physch. Rev. 85, 582 (1978).

Wood, C., "Implications of simulated lesion experiments for the interpretation of lesions in real nervous systems", in Neural Models of Language Processes, ed. by M. A. Arbib, D. Caplan and J. C. Marshall, (Academic Press, New York) (in press).

Zucker, S. W., Y. G. Leclere and J. Mohammed, "Continuous relaxation and local maxima selection-conditiosn for equivalence", IEEE Trans. on Pattern Analysis and Machine Intelligence, PAMI-3 117-127 (1981).